CITE-On: a deep learning approach for online cell identification and trace extraction in functional two-photon calcium imaging

Luca Sità[1,*,#], Marco Brondi[1,*,#], Pedro Lagomarsino de Leon Roig [1,2], Sebastiano Curreli[1], Mariangela Panniello[1], Tommaso Fellin[1,#]

[1] Optical Approaches to Brain Function Laboratory, Istituto Italiano di Tecnologia, Genova, Italy

[2] University of Genova, Genova, Italy

* Equal contribution

Corresponding authors:

Luca Sità, Optical Approaches to Brain Function Laboratory, Istituto Italiano di Tecnologia, Via Morego 30, 16163 Genova, Italy, tel: +39 010 2896549, fax: +39 010 2896230, email: luca.sita@iit.it

Marco Brondi, Optical Approaches to Brain Function Laboratory, Istituto Italiano di Tecnologia, Via Morego 30, 16163 Genova, Italy, tel: +39 010 2896549, fax: +39 010 2896230, email: marco.brondi@iit.it

Tommaso Fellin, Optical Approaches to Brain Function Laboratory, Istituto Italiano di Tecnologia, Via Morego 30, 16163 Genova, Italy, tel: +39 010 2896549, fax: +39 010 2896230, email: tommaso.fellin@iit.it

Keywords: two-photon functional imaging, genetically encoded calcium indicators (GECI), deep neuronal networks, online image segmentation

Abstract

In vivo two-photon calcium imaging is a powerful approach in experimental neuroscience. However, processing two-photon calcium imaging data is computationally intensive and time-consuming, making online frame-by-frame analysis a challenge. This is especially true for large field-of-view (FOV) imaging. Here, we present CITE-On (Cell Identification and Trace Extraction Online), a convolutional neural network-based algorithm for fast automatic cell identification, segmentation, identity tracking, and trace extraction in two-photon calcium imaging data. CITE-On processes thousands of cells online, including data from mesoscopic two-photon imaging, and provides accurate functional measurements from most neurons in the FOV. Applied to publicly available datasets, CITE-On achieves performance similar to that of state-of-the-art methods for offline analysis. Moreover, CITE-On generalizes across calcium indicators, brain regions, and acquisition parameters in anesthetized and awake head-fixed mice. CITE-On represents a powerful new tool to speed up image analysis and facilitate closed-loop approaches, for example in combined all-optical imaging and manipulation experiments.

Introduction

Multi-photon imaging in combination with Genetically Encoded Calcium Indicators (GECI) allows the recording of population activity with high spatial resolution in the intact brain *in vivo* [1-6]. However, multi-photon imaging datasets, in the form of t-series, can be heavy (0.5 GB to > 1 TB) and their processing requires time and computational power. More specifically, the precise identification and segmentation of neuronal structures (typically somata) in a given FOV is critical to extract truthful and reliable information from raw imaging t-series [6]. This step can be complex because of dense GECI staining, low signal-to-noise ratio (SNR), presence of motion artifacts, and large number of neurons in the FOV (e.g., in the case of mesoscopic two-photon imaging [7, 8]).

Segmentation is typically performed in two ways: *i)* manually, based on visual inspection by an expert user and on selection of pixels into regions of interest (ROIs); *ii)* automatically, employing supervised or unsupervised methods leveraging on static and dynamic properties of the fluorescence signal in the t-series [9-23]. Manual segmentation [24, 25] is time consuming and impractical in case of large datasets and FOVs (e.g. mesoscopic imaging) or when real-time manipulation of experimental conditions is needed [26, 27]. State-of-the-art automatic approaches apply pixel correlation [15, 16, 19], principal/independent component analysis (PCA/ICA) [15, 16], constrained non-negative matrix factorization (CNMF) [10, 14, 17], and deep neural networks (DNN) [10, 20-22] to perform FOV segmentation. These approaches are usually applied offline and generally take advantage of both neuronal spatial footprints and the temporal dynamics of the fluorescence signal associated to the identified spatial footprints. Consequently, their performance benefits from long acquisitions [10, 16, 23], with highly active cells being more easily segmented than rarely active or inactive ones [10, 22]. Moreover, current methods often require the experimenter to set initialization parameters ahead of the segmentation process [10, 11, 15-17, 22]. While most of these parameters are generally easy to adjust (*e.g.* frame rate and indicator kinetics), some are inaccessible to the user online (*e.g.* number of expected ROIs in a FOV and spatial constraints on ROI shapes) and must be determined through multiple offline rounds of empirical tuning steps.

The quality of *in vivo* two-photon calcium imaging is also extremely sensitive to motion artifacts [6]. In particular, the shape and position of imaged cells may change due to motion artifacts correlated with the animal's locomotion, breathing, and heartbeat. In current approaches [10, 15-17, 22, 28], successful neuronal segmentation is typically achieved after correcting for motion artifacts: a process requiring additional time and computational power. The output of the segmentation process is thus a static

mask, representing the "average" shape and position of each cell throughout the t-series. This approach is impractical whenever cells should be tracked online on a frame-by-frame basis, for instance when a neuronal ensemble (i.e. a group of coactive neurons) must be optogenetically manipulated after being identified [29, 30]. In fact, neuronal ensembles are dynamic and different cells may belong to a given ensemble at a certain time instant, making it difficult to define *a priori* the neuronal identities belonging to future ensembles [31]. Finally, downstream of segmentation, the dynamic fluorescence signal from each cell must be extracted and "decontaminated" from background signal [6, 32]. Different approaches are available to this end [10, 11, 15-17, 32, 32, 33], all requiring additional computational time. As a result of all these analytical steps, a total processing time of 30 to 90 minutes was reported for most efficient methods when processing FOVs of about 500 x 500 μm^2 containing hundreds of cells imaged over tens of thousands frames [10, 11, 22]. Altogether, current analytical approaches are: *i)* still limited in their ability to perform online analysis, which is necessary for closed-loop experiments; *ii)* biased against the identification of rarely active or inactive cells, which could be as informative to target as more active neurons, for example, in longitudinal all-optical imaging and manipulation approaches; *iii)* not validated on large FOVs, such as those generated by mesoscopic imaging.

Here, we describe CITE-On, a CNN-based algorithm trained to perform neuronal somata identification in two-photon imaging recordings, combined with a light-weight dynamic segmentation and trace extraction pipeline. CITE-On identified hundreds to thousands of neuronal cell bodies on a frame-by-frame basis. Moreover, CITE-On also identified inactive neurons, removing biases towards highly active cells. CITE-On's light architecture and processing strategy allowed, for the first time, fast automatic segmentation, tracking, and trace extraction in mesoscopic two-photon imaging t-series.

Results

CITE-On: structure and analysis pipeline

CITE-On accepted individual frames from two-photon calcium imaging t-series (Fig. 1a) and it comprised two main parts: an image detector based on the publicly available convolutional neural network (CNN) RetinaNet [34] dedicated to the identification of neuronal somata and a custom-built downstream light-weight analysis pipeline, designed for functional trace extraction. The image detector and the analysis pipeline operate as asynchronous parallel processes, in order to provide discrete cell detection update (up to 10 Hz, see text below) and faster than real time functional traces (available at 100 Hz under all experimental conditions tested in this study, see text below). CITE-On required three pre-processing steps ahead of the CNN image detection: *i)* frame downsampling; *ii)* image upscaling; *iii)* replication of the input image into three identical channels (Fig. 1a). The frame downsampling value was set according to the image SNR, while the upscaling factor depended on the ratio between the FOV surface and the average surface of the neuronal somata (see Methods and text below). The length of the frame downsampling window, as well as the value of the upscaling factor, was defined *a priori* and adjusted according to the data to maximize performance. The three identical images were sent to the CNN for image detection (blue rectangle in Fig. 1). The output of the CNN (yellow highlights in Fig. 1) was a set of boxes tightly surrounding each detected cell soma ("bounding boxes", represented as green squares over the FOVs in Fig. 1a-d). Coordinates and surfaces of each bounding box were used in the analysis pipeline, to generate: *i)* cell identity assignment and tracking along the t-series; *ii)* dynamic segmentation of neuronal somata; *iii)* background subtraction procedure and extraction of neuronal functional trace.

CITE-On worked both offline, after the acquisition was completed and the whole t-series was available (Fig. 1b), or online, using individual imaging frames as inputs, continuously streamed from the experimental set-up during the acquisition of the imaging t-series (Fig. 1c-d). In the offline pipeline, a Fourier-transform approach [35] was used to correct for planar motion artifacts throughout the t-series (Fig. 1b). The frame downsampling corresponded to the projection of the entire t-series onto its temporal median calculated across all frames (Fig. 1b). Soma detection was then performed once on the pre-processed median image (detection in Fig. 1b), and bounding boxes were generated for each frame of the t-series (Fig. 1b, yellow highlights). Each bounding box was associated with a score, representing network confidence in cell detection. Bounding boxes with intersection over union (IoU) < 20 % were considered as separate neuronal identities. When IoU of two bounding boxes was > 20 %, the bounding box with the highest score was retained.

In the online pipeline, no motion correction was performed, and the user selected between two downsampling strategies depending on the SNR of the data, and on the required upscaling factor in the pre-processing step. In case of relatively high SNR and low upscaling factors (Fig. 1c), a sliding average was calculated on the first *n* frames of the t-series and updated with every new individual frame generated by the microscope. Neuronal detections were updated for each imaging frame starting from the $n+1^{th}$ frame. When the SNR was relatively low and the upscaling factor large (Fig. 1d), a step average approach was performed, where the input for the image detector was the average projection of blocks of *n* frames. Additional *n* frames were thus required for generating the next step average projection and the detections were updated every *n* frames. The maximum detection rate decreased with the upscaling factor, with a peak rate of 10 Hz with upscaling factor equal to 1 (Fig. 1e). Active detections (*i.e.* detections in the current sliding average or step average) and past detections (*i.e.* detections in any previous sliding or step average) were continuously tracked and

updated (Fig. 1c-d, detection update and I.D. tracking). Specifically, active detections were compared with past detections at each step of detection update and a new identity was added (and included in the tracking system) every time the surface of an actively detected bounding box had IoU < 25 % with any of the previously identified boxes. Bounding boxes from active detections with IoU > 25 % with those of past detections did not change identities of previously detected boxes, but their positions and shapes were updated according to the position in the most recent detection step. All past detections without updates were retained in the tracking system in the form of their last active detection for the remaining part of the t-series (Supplementary Movie 1).

For both the online and offline pipelines (Fig. 1f-g), bounding boxes were used to generate a dynamic segmentation of the t-series and to identify ROIs. The distribution of fluorescence values inside each bounding box was computed at each frame (Fig. 1f, left). Only pixels with values between the 80^{th} and the 95^{th} percentile of the box's fluorescence distribution were assigned to the ROI corresponding to the cell soma (white pixels of the binary mask in Fig. 1f, right). Since pixel assignment to cell somata in each individual box was updated at each frame, the resulting dynamic segmentation was updated online for every new frame. All the FOV pixels that were not included in any bounding box were assigned to a global background ROI. The fluorescence intensity of all pixels belonging to the global background ROI was averaged at each frame to obtain the background signal (*bg*). Moreover, at each frame, the *bg* was subtracted from the fluorescence of each segmented neuronal ROI, generating functional fluorescence traces (Fig. 1g). Since shape, number, and position of bounding boxes changed as the t-series progressed (according to active detections and tracking), the pixels assigned to *bg* also changed in number and identity across frames. Identity tracking, segmentation, and functional trace extraction required on average 10 ms *per* frame (either offline or online).

Ground truth generation and training of the image detector

The ResNet50 Feature Extractor CNN incorporated in CITE-On was not originally developed for detecting neuronal somata, rather for the analysis of natural images and it was trained on > 1 million RGB images across 80 classes (http://www.image-net.org/). We decided to use a transfer learning strategy [36] to adapt this efficient detection architecture to the identification of neuronal somata (*i.e.* a single class) in grey-scale two-photon images. This choice was dictated by the fact that available two-photon calcium imaging datasets (http://neurofinder.codeneuro.org/, http://help.brain-map.org/display/observatory/Data+-+Visual+Coding) are far too small for an *ab-initio* CNN training. Moreover, they are too homogeneous in terms of calcium indicator used, FOV dimensions, cell density, acquisition frame period, SNR and background signal contamination [22], making them suboptimal even for a transfer learning strategy. For example, no publicly available large dataset comprises imaging data collected using red-shifted GECIs, such as jRCaMP1a. We thus decided to use a dedicated dataset for training and internal validation. In this way, we employed publicly available datasets to test CITE-On performance and its generalization capability on never-before-seen data. The dedicated dataset included 197 t-series from 28 mice acquired using different acquisition parameters (see Methods). More specifically, we included 121 t-series from layer IV neurons of the somatosensory cortex expressing either GCaMP6f, GCaMP6s, or GCaMP7f (globally indicated as "LIV") and 76 t-series from the CA1 pyramidal neurons of the hippocampus expressing both GCaMP6f and jRCaMP1a (indicated as "CA1 GCaMP6f" and "CA1 jRCaMP1a", respectively). We included t-series with heterogeneous median fluorescence and SNR in order to reduce potential biases toward bright cells during the training process, while avoiding large differences between groups of data that could have generated better performance on specific subsets of t-series (Supplementary Fig. 1a, b).

To obtain a consensus Ground Truth (GT) annotation of the t-series used for training and validation of the CNN, two human graders manually annotated all t-series, defining the tightest rectangular box fitting each visible cell in each FOV. Manual GT annotation was preferred to automatic segmentation for two main reasons: *i)* available automatic segmentation approaches rely on both functional (i.e. fluorescence signal dynamics across frames) and morphological features [9-23], while we wanted the GT annotation to be exclusively based on morphological features (see below); *ii)* manual annotation is still frequently considered more accurate than automatic methods [10, 22]. Initially, manual annotation on single frames by two graders produced only few neuronal identities because cells were only visible in a minority of frames (Supplementary Fig. 1c, Supplementary Movie 2). This could be due to the low basal emission of some of the indicators used (e.g. GCaMP6f), to the variable expression level of the calcium indicator across cells, and to the sparse activity profile of the imaged cells. In order to increase the visibility of neurons, we created high contrast single images representative of each t-series. To this end, we first corrected each t-series for planar motion artifacts, and then collapsed each acquisition onto its median projection (Supplementary Fig. 1d). These images were sharpened (Supplementary Fig. 1e) and gamma corrected. Brightness and contrast were adjusted in order to obtain a distribution of intensity values spanning the whole bit range. The sharpened images, named enhanced median projection (EMP) (Supplementary Fig. 1f, g), were used for manual annotation (Supplementary Fig. 1h, Supplementary Movie 3). In training and validation datasets, grader #1 annotated 14,425 boxes, while grader 2 annotated 12,912 (Supplementary Table 1). The bounding boxes produced by grader #1 and grader #2 and their superposition in different experimental preparations are shown in Supplementary Fig. 2. Annotations from the two graders largely overlapped (weighted average precision or mAP, 0.77 ± 0.08; F-1 score, 0.93 ± 0.02; precision: 0.98 ± 0.01; recall: 0.88 ± 0.12, N = 197 EMPs, Supplementary Table 2). Given the high similarity of the independent annotations provided by the two graders, we defined the consensus GT for our entire

dataset as the union of the two sets of annotations. Given that our dataset contained partially overlapping FOVs and more than one t-series acquired on the same FOV, the dataset was manually split into training (160 t-series) and validation (37 t-series) subsets. To avoid data leakage and to decrease overfitting [37], t-series from the same or similar FOVs were included in either the training dataset or the validation dataset. We trained the CITE-On image detector on the training dataset and evaluated its performance on the validation dataset.

Performance of the image detector

We trained the CITE-On image detector on our consensus GT annotations achieving the best performance after 17 training epochs (mAP: 0.79 on the validation dataset). We first evaluated CITE-On performance using the offline pipeline on the validation dataset. A representative CITE-On output for a CA1-GCaMP6f, a CA1-jRCaMP1a, and a LIV t-series is shown in Fig. 2a-d and Fig. 2e, respectively. For the whole validation dataset, the mean ± s.d. F-1 score, Precision, and Recall are reported in Fig. 2f and Supplementary Table 3.

We then used the online pipeline and calculated the F-1 of CITE-On detections on the motion corrected validation dataset, which was used as input to CITE-On at the actual frame rate occurring during acquisition. Our validation data required an upscaling factor of 1, compatible with a maximum detection rate of 10 Hz, while acquisition frame rates varied between 1.5 Hz for LIV and 3 Hz for CA1 acquisitions. We empirically explored the effect of frame downsampling on detection performance using the sliding average approach with different numbers of frames (*n*). Our aim was to maximize F-1 and score threshold for detections, while minimizing *n*. F-1 increased with *n* between 1 and 20 frames, value at which the absolute maximum was observed (Fig. 3a). Using the sliding average, an initial delay of 6.6 s for LIV data and 14 s for CA1 data was necessary before CITE-On

processed each frame in real time at a detection rate of 10 Hz. F-1 values calculated on sliding averages of 20 frames are reported in Fig. 3a. Detections (green boxes) are displayed together with GT annotations (red boxes) for both the GCaMP6s (Fig. 3b, left) and the jRCaMP1a channels (Fig. 3b, middle). The superimposition of the detections from both channels is shown on Fig. 3b (right). Similarly, in Fig. 3c, we show GT and online detections (red and green boxes, respectively) for a representative LIV (GCaMP6s) t-series.

To quantify the impact of motion artifacts on online detection accuracy, we calculated the F-1 score on validation t-series that were not corrected for motion artifacts (Supplementary Table 4). To this end, we translated the GT annotations for each frame according to the shift vectors produced by the motion correction algorithm implemented when building the relative EMPs (see Methods). We observed no significant difference between the F-1 obtained on motion corrected and non-corrected validation t-series (Fig. 3d), suggesting that the motion correction step was not necessary to achieve higher performance with our approach. The F-1 values across frames for the non-motion-corrected data are reported in Fig. 3e. The distribution of motion displacements is shown in Supplementary Fig. 3.

Online data processing

We developed a fast method to dynamically segment each detected cell based on the corresponding bounding box, relying on the instantaneous (i.e. frame-wise) fluorescence statistics of the pixels inside each box (Fig. 4, Supplementary Movie 4). Specifically, the fluorescence intensity distribution of the pixels inside each bounding box was first computed frame-wise. Pixels with fluorescence values between the 80^{th} and 95^{th} percentile of the distribution were then selected as belonging to neuronal somata. The values of selected pixels were averaged, and the resulting fluorescence trace

was "denoised" by subtracting, at each frame, the *bg* signal. This simple method was computationally light, an important requirement to achieve fast frame-by-frame data processing (trace extraction rate, 100 Hz). Bounding boxes detected by CITE-On on a representative LIV t-series and a representative CA1 t-series are shown in Fig. 4a. Representative fluorescence traces extracted by CITE-On on the two t-series are displayed in Fig. 4b-c. Fig. 4d-e shows the cross correlation matrix (lower left triangle) and the dendrogram analysis (upper right triangle) of all the identified cells before (Fig. 4d-e left) and after (Fig. 4d-e, right) *bg* subtraction.

We compared traces extracted by CITE-On with those extracted by CaImAn, a state-of-the-art method based on CNMF [10]. We used the bounding boxes generated offline by CITE-On to build binary masks that were used as seeds to initialize the seeded-CNMF algorithm [10]. The spatial components of the CNMF were non-zero only inside the bounding boxes identified by CITE-On, therefore, the detected factors from seeded-CNMF had one-to-one correspondence with the detected boxes from CITE-On. We first observed very high pairwise cross correlations between the *bg* traces extracted with the two methods (Supplementary Fig. 4a). We then asked how the *bg* signal calculated over the whole FOV (*bg*) correlated with the 'local *bg*', that is, the background activity calculated for each cell from the pixels in the immediate surroundings of the relative bounding box (see Methods). The average correlation between *bg* and local *bg* traces was high (Supplementary Fig. 4b). Given the high correlation values observed between *bg* and local *bg* and the lower computational cost of *bg*, we decided to implement the *bg* method only. We then tested how the *bg*-subtracted functional traces calculated by CITE-On compared with those extracted with seeded-CNMF and we again observed high correlation values (Supplementary Fig. 4c).

Although the average correlation values for the cells extracted with the two methods were high, some neuronal pairs showed lower correlations. We asked whether the low correlation values emerged from pairs of cells with low SNR. We found that pairwise correlation values for traces extracted with the two methods increased with the SNR of the corresponding neuronal traces (Supplementary Fig. 4f), indicating that indeed the functional traces obtained with the two methods were more similar when the trace SNR was high.

Offline performance on never-before-seen recordings

To test the robustness of our image detection approach and its ability to generalize across experimental conditions, we tested CITE-On on three additional datasets, which were not used during the training and validation phases. The three datasets were: the Allen Brain Observatory repository (ABO, 19 t-series divided into 9 superficial, ABO_{sup}, t-series, acquired in visual cortex at depth 175 µm, and 10 deep, ABO_{deep}, t-series, acquired in visual cortex at depth 275 µm), the Neurofinder Challenge dataset (28 t-series, divided into 19 t-series, NF_{train}, 9 t-series, NF_{test}, from different preparations at different depths), and a dataset acquired in our laboratory using GRIN-based endoscopic two-photon imaging of the ventral posteromedial thalamic nucleus (VPM, 9 t-series). The three datasets were first manually annotated *de novo* to obtain the GT annotation (Supplementary Fig. 5, Supplementary Table 1-2). Because the ratios between FOV and cell surface were variable across the ABO, NF, and VPM datasets, and different from our validation dataset, we optimized the upscaling factor and used the one that maximized the F-1 score (Fig. 5a-e). The offline performance (Precision, Recall and F-1 score) obtained using optimized upscaling factors for each dataset is shown in Fig. 5f and Supplementary Table 3.

We compared the offline detection performance of CITE-On (Supplementary Table 3) with state-of-the-art alternative segmentation approaches such as STNeuroNET [22], $CaImAn_{online}$ [10], $CaImAn_{batch}$ [10], Suite2P [16], HNCcorr [19], UNet2DS [23] on the ABO and NF datasets provided in [22]. We used our GT annotations to test CITE-On and those reported in [22] to test the alternative methods. We found that the average CITE-On performance was equivalent to that computed with other methods (Fig. 6). Moreover, the background signals calculated using seeded CNMF and CITE-On presented high cross correlation values for all three datasets (Supplementary fig. 6a). Similarly, high cross correlation was measured between local *bg* and *bg* signals (Supplementary Fig. 6b), as well as between CITE-On extracted functional traces after global *bg* subtraction and CaImAn extracted traces (Supplementary Fig. 6c). As for our previous characterization on the validation dataset, cross correlation values of the extracted traces with CITE-On and CaImAn increased with SNR in all datasets (Supplementary Fig. 6f).

We compared true positive detections obtained with CITE-On with ABO true positive detections and with STNeuroNET true positive detections (Fig. 7a-c). Examples of cells identified by CITE-On and present in the true positive detections from both ABO and STNeuroNET are reported in Fig. 7 d, together with their relative background-subtracted traces extracted by CITE-On. We analyzed the functional traces of all the CITE-On true positives (i.e. including cells detected in ABO and STNeuroNET) after background subtraction (Fig. 7f). Similarly to what described for our validation dataset, the dendrogram sorting showed blocks with different cross correlation values for various subgroups of cells. On average, the number of CITE-On detected cells exceeded the number of identities available in public repositories (Supplementary Fig. 7, Supplementary Table 1).

Since state-of-the-art segmentation methods have a bias against inactive cells, we investigated whether CITE-On-only cells were inactive. We quantified the number of calcium events *per* detection in the ABO true positives (ABO TP), STNeuroNET true positives (STNuroNET TP), CITE-On true positives (CITE-On TP), and CITE-On-only true positives (CITE-On-only, Supplementary Fig. 7). The number of cells with few detected calcium events was larger for CITE-On TP (Supplementary Fig. 7). Moreover, although some of the CITE-On-only cells were silent (as expected), a large fraction of them displayed detectable activity (in the whole ABO dataset, 91 % displayed at least one calcium event and 69 % showed at least ten calcium events, Supplementary Fig. 7a-d). Moreover, the structure of the dendrogram built with CITE-On-only cells resulted similar to the one obtained with the complete set of CITE-On true positives (Fig. 7g), indicating that the group of active cells detected exclusively by CITE-On recapitulated the functional structure of most active neurons. The number of detected calcium events with CITE-On (all detections, true positives, false positives, and CITE-On-only identities), ABO (true and false positives), and with STNeuroNET (true and false positives) is reported in Supplementary Fig. 7e. The number of detections in the different datasets is reported in Supplementary Table 5.

Online performance on never-before-seen recordings

We ran CITE-On online using our GT annotation on each frame of the ABO, NF, and VPM datasets. When computing the F-1 score, we tested different sizes of the sliding average, in order to define the smallest number of frames required to achieve real-time processing. The absolute maximum F-1 score was achieved in 10 frames for the ABO dataset (both ABO_{sup}, and ABO_{deep}), 20 frames for the VPM dataset, and 200 frames for the NF (train and test) dataset (Fig. 8a-e). For the ABO dataset, 0.3 s were necessary to acquire 10 frames (0.00086 % of the whole time series of average duration 115,635 ± 130 frames) with a CITE-On detection rate of 5 Hz and an upscaling factor of 2. The NF dataset

358 required upscaling factors of 2.4 and 2.6 for NF_{train} and NF_{test}, respectively. A time window of 28.5 s
359 was required to acquire the 200 frames necessary to reach peak F-1 value (5.4 % of the whole t-series
360 of average duration 3,697 ± 1,874 frames) with a CITE-On detection rate of 4 Hz for NF_{train} and 3 Hz
361 for NF_{test}. Given the relatively low online performance for these latter datasets, we decided to measure
362 the F-1 score using an alternative frame downsampling strategy: step average. Using this approach,
363 we found that detection performance remained high (Fig. 8i and Supplementary Table 4). The F-1
364 score calculated online as a function of the length of the t-series for ABO (ABO_{sup} and ABO_{deep}
365 together), NF (NF_{train} and NF_{test}) and VPM data is shown in Fig. 8j. Stable F-1 scores were observed
366 within 30 % of the total length of the t-series. Processing time required for running the online pipeline
367 on the ABO, VPM, and NF datasets was not different from what described previously for our
368 validation datasets.
369
370 *Analysis of large FOV mesoscopic images*
371 Given the speed and light-weight of the CITE-On architecture, we tested if it could be applied for
372 detecting cells in the mesoscopic imaging t-series described in [7]. Because the dimensions of the input
373 image were too large (1792 pixels x 1682 pixels, or 4.8 mm x 4.8 mm, 1 μm pixel size) to fit the CNN
374 architecture using the appropriate upscaling factor based on the FOV/neuron surface ratio (upscaling
375 factor for direct processing, 12.8), we tiled the entire mesoscopic FOV in 272 subfields (subfield
376 dimension, 128 pixels x 128 pixels, 28 pixels overlap). Each subfield was appropriately upscaled
377 (upscaling factor, 1; score threshold, 0.4; detection rate, 10 Hz). To increase speed, we multiplexed
378 the CNN detector process and processed subfields in batches of 8 images in parallel, until completion.
379 Identity duplicates were suppressed using a non-maximum suppression algorithm, where boxes
380 having an IoU > 25% were considered duplicates and only the one with the highest score was retained
381 (see Methods). Image detection outputs were finally recombined to reconstruct the entire FOV. In

Fig. 9a, we report CITE-On detected bounding boxes on the entire FOV obtained using the offline pipeline. Two FOV patches are magnified (Fig. 9b) to show the shape of the detected cells (total number of detected somata, 4,842). In Fig. 9c, we show representative fluorescence traces obtained with CITE-On from five cells from the whole mesoscopic FOV. Fig. 9d shows the cross correlation matrix (lower left triangle) of all the identified cells for the first 700 frames of the t-series. The dendrogram analysis (upper right triangle of Fig. 9d) highlighted several distinct functional modules observed in the identified neuronal population.

The single subfields were processed by the CITE-On detector at 10 Hz (upscaling factor, 1), while segmentation, tracking, and functional trace extraction were performed at 100 Hz (see previous results). Parallel processing of all the 272 sub-fields required 12.6 s for each detection step. Therefore, with a step average downsampling approach including 25 frames (13.2 s of mesoscopic imaging time since acquisition rate was 1.9 Hz), and while extracting traces faster than the incoming frames, we minimized the CITE-On detection lag with respect to the running acquisition. With this strategy, we achieved an online F-1 score of 0.54 with a score threshold of 0.25 (quantified on four patches from the entire FOV). CITE-On can thus be efficiently applied for fast processing of mesoscopic two-photon t-series.

Discussion

In this study, we developed a CNN-based algorithm, CITE-On, for fast analysis of two-photon imaging recordings. CITE-On performed online accurate identification of neuronal somata, tracking of identities across frames, dynamic segmentation, and functional trace extraction with background subtraction. CITE-On generalized across calcium indicators, brain regions, acquisition parameters, and it was successfully applied to data obtained using different surgical and optical preparations (e.g.

chronic superficial imaging window, chronic deep imaging window, and endoscopic GRIN lens-based deep imaging).

Our image detection strategy was based on RetinaNET, a CNN originally developed to detect natural images [34]. On one side, this choice was justified by RetinaNET's excellent performance in object recognition and by its readiness. On the other hand, it required us to exploit RetinaNET on a set of images, grey-scale two-photon fluorescence images of neurons, that were remarkably different from those RetinaNET was originally trained on. To compensate for this difference, and to have a large and heterogeneous dataset for training and validation of the detection algorithm, we built a dedicated library of hundreds of two-photon imaging t-series acquired with different GECIs, in different regions of the mouse brain, at different frame rates, using different surgical/optical preparations, and showing variable image quality. In this dataset, a reliable GT consensus was reached using the annotations of two human graders, which allowed us to evaluate CITE-On performance. To obtain a GT annotation insensitive to the graders' biases, a potential alternative approach could have been to generate an *in silico* dataset for network training and validation [38]. No online libraries of this kind are currently available, but we foresee that this approach may represent an extremely useful method to optimize future CNN-based approaches for the analysis of two-photon functional data. It is also worth noticing that utilizing public datasets already used for training and validation of alternative processing toolboxes [22] would have given us the possibility to take advantage of third-party GT annotations. However, we decided not to do so because: *i)* CITE-On would have likely inherited the annotation bias toward more active cells, which is shared by existing publicly available repositories; *ii)* by using public datasets exclusively for the validation of the CITE-On image detector (rather than for training, too), we avoided any chance of data leakage, and we demonstrated that CITE-On generalized well to never-before-seen data.

CITE-On performed as state-of-the-art algorithms [10, 11, 16, 19, 22, 39] on publicly available datasets and, importantly, it did so in a much shorter time. In fact, only a few seconds were needed to have online, frame-by-frame, accurate ROI segmentation, identity tracking, *bg* subtraction, and functional trace extraction. Four main characteristics were crucial for CITE-On's high performance. First, CITE-On relied exclusively on morphological features to identify neurons. Second, neuronal identification was dynamic and it adapted to changes in shape, position, and activity of the detected cells frame-by-frame, avoiding time-expensive motion correction procedures. Third, once bounding boxes were identified in individual frames, we used a simple cost-effective strategy to extract pixels belonging to neuronal ROIs based on pixel's brightness. Fourth, we implemented a fast background subtraction strategy, limiting computational costs. When applied in the online modality, these characteristics were crucial to achieve real-time frame-by-frame trace extraction, something current approaches do not achieve [9-25], while maintaining high cell detection performance.

The most common automatic analytic pipelines for the analysis of two-photon imaging t-series typically rely on both spatial features and fluorescence signal fluctuations [9-23]. This necessarily requires the use of several, if not all, frames in a t-series, precluding efficient online analysis. In contrast, CITE-On was conceived such that the analysis pipeline was performed frame-by-frame based on morphological features. This dynamic analytical process was continuously updated over time allowing the algorithm to cope with planar motion artifacts, a feature that, unlike other available methods, enabled us to bypass the motion correction pre-processing step and to save further computational time. Importantly, neuronal identities identified frame-by-frame were reliably tracked over the t-series. To extract pixels belonging to neuronal ROIs within a bounding box, we used a criterion based on pixel brightness and to correct for background contamination we devised a global

rather than local neuropil computation. These procedures had the advantage of being simply and simultaneously implementable for thousands of cells, allowing the extraction of functional traces at low computation cost. The observation that functional fluorescence traces extracted by CITE-On were highly correlated with those extracted on the same bounding boxes by a state-of-the-art method, i.e. CaImAn [10], confirmed the validity of our computationally effective approach.

Thanks to the features described above, CITE-On efficiently processed full mesoscopic two-photon t-series (FOV dimension, 5 mm x 5 mm). CITE-On analyzed full mesoscopic images dividing each image in subfields and processing subfields in parallel. CITE-On's detector processed single subfields at 10 Hz, while segmentation, tracking, and functional trace extraction were performed at 100 Hz. Parallel processing of all the 272 sub-fields generating a whole mesoscopic FOV required 12.6 s for each detection step. Thus after 12.6 s, trace extraction could be performed at 100 Hz on thousands cells. Besides it application online, the offline application of CITE-On is also going to be extremely powerful for the identification of the thousands neurons imaged in mesoscopic two-photon functional imaging.

Closed-loop all-optical experiments are fundamental to investigate whether models of network dynamics, circuit connectivity, and causality are accurate [26]. Recently, all-optical closed-loop experiments have been validated [27]. For example, using this approach specific groups of neurons were activated based on the readout of ongoing activity in a reference cell. However, the closed-loop strategy described in [27] was based on *a priori* identification of the reference cell. Because CITE-On allows efficient frame-by-frame cell identification and trace extraction, it will enable a new type of experiment in which the loop is closed based on real time identification and readout of any neuron or group of neurons in the FOV.

CITE-On differs in several main features from OnACID [11], a recent method that combines CNMF and fast deconvolution for the online analysis of calcium data. First, CITE-On does not need any initialization step in the online pipeline. In contrast, OnACID requires the user to provide the expected number of cells to be identified, and it starts the online analysis based on the CNMF output of an initial portion of the t-series (typically 1000-3000 frames), generating an initial temporal lag of 2-4 minutes [11]. Second, CITE-On does not require the correction of motion artifacts in the online pipeline, saving further computational time. Third, CITE-On performs tracking, dynamic segmentation, and functional trace extraction processes at 100 Hz independently of the number of detected neurons and their activity. This feature allows maintaining high online performance on FOVs characterized by large number of neurons (e.g., those obtained with mesoscopic two-photon imaging) and sparse activity. Four, CITE-On does not use local pixel correlation for cell identification, which may be advantageous to separate nearby synchronous cells. In contrast, the OnACID fast deconvolution approach may be more efficient in separating adjacent cells with different temporal profiles of fluorescence emission [11]. Finally, OnACID was tested on two datasets with rather homogenous acquisition parameters [11]. Here, we demonstrate that CITE-On generalizes across indicators (e.g., GCaMP6s, GCaMP6f, GCaMP7f, and jRCaMP1a) and across data acquired with different pixel size, SNR, and frame rate. This property adds flexibility in the application of online analysis tools to different experimental conditions.

Because cell identification was based only on spatial features, CITE-On identified both active and silent cells. This unique characteristic of CITE-On is important because it adds further flexibility in designing imaging experiments. Neurons that are silent in a t-series may change their level of activity in subsequent acquisitions depending on the behavioral state of the animal or because of external

manipulations [40-42]. Thus, being able to track cells regardless of their activity level is key, for instance, for investigating the sensory information beard by neurons that dramatically change their activity throughout longitudinal imaging experiments. Biasing the analysis toward active neurons, as currently done by most approaches, intrinsically skews the proportion of cells that are responsive to a given stimulation in a certain brain region. In this regard, it is also interesting to note that neurons that were detected only by CITE-On and not by other state-of-the-art approaches comprised silent cells but, unexpectedly, also active cells. Active CITE-On only cells may be associated with low SNR of fluorescence signals and they may be missed by current approaches [10].

In summary, we developed CITE-On, a new tool to effectively process two-photon imaging data frame-by-frame, while maintaining similar cell detection and trace extraction performance of existing offline state-of-the-art methods. Future developments of CITE-On will likely include its optimization for one-photon imaging [43, 44], its application to genetically encoded voltage indicators [45] as well as to volumetric two-photon imaging [46].

Author contributions

LS, MB, PL developed software and performed analysis. MB, SC performed experiments. LS, MB, TF conceived the project. TF coordinated the project. LS, MB, MP, TF wrote the paper with inputs from other authors. All authors approved the final version of the manuscript.

Acknowledgments

We thank Dr. C. Arlt and Dr. C. Harvey for sharing mesoscopic imaging data, Dr. A. Sattin for sharing VPM recordings, Dr. S. Fiorini and Dr. A. Barla for discussion and comments on algorithm development, Dr. D.S. Kim and the GENIE project and Dr. J.M. Wilson for constructs, Dr. S. Succol

for technical support, and Dr. D. Vecchia for help with the figures. This work was supported by the European Research Council (ERC, NEURO-PATTERNS) and NIH Brain Initiative (U19 NS107464).

Declaration of interest

The authors declare no competing interests.

References

1. Helmchen,F. & Denk,W. Deep tissue two-photon microscopy. *Nat Methods* 2, 932-940 (2005).
2. Wang,T. *et al.* Three-photon imaging of mouse brain structure and function through the intact skull. *Nat. Methods* 15, 789-792 (2018).
3. Chen,T.W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* 499, 295-300 (2013).
4. Dana,H. *et al.* Sensitive red protein calcium indicators for imaging neural activity. *Elife.* 5, 12727 (2016).
5. Svoboda,K. & Yasuda,R. Principles of two-photon excitation microscopy and its applications to neuroscience. *Neuron* 50, 823-839 (2006).
6. Harris,K.D., Quiroga,R.Q., Freeman,J., & Smith,S.L. Improving data quality in neuronal population recordings. *Nat Neurosci.* 19, 1165-1174 (2016).
7. Sofroniew,N.J., Flickinger,D., King,J., & Svoboda,K. A large field of view two-photon mesoscope with subcellular resolution for in vivo imaging. *Elife.* 5, (2016).
8. Tsai,P.S. *et al.* Ultra-large field-of-view two-photon microscopy. *Opt. Express* 23, 13833-13847 (2015).

9. Diego,F. & Hamprecht,F. Sparse space-time deconvolution for Calcium image analysis. *Prooceedings of the 27th International Conference on Neural Information Processing Systems* 1, 64-72 (2014).

10. Giovannucci,A. *et al.* CaImAn an open source tool for scalable calcium imaging data analysis. *Elife.* 8, (2019).

11. Giovannucci,A. *et al.* OnACID: Online analysis of calcium imaging data in real time. *Advances in Neural Information Processing Systems*(2017).

12. Guan,J. *et al.* NeuroSeg: automated cell detection and segmentation for in vivo two-photon Ca(2+) imaging data. *Brain Struct. Funct.* 223, 519-533 (2018).

13. Kaifosh,P., Zaremba,J.D., Danielson,N.B., & Losonczy,A. SIMA: Python software for analysis of dynamic fluorescence imaging data. *Front Neuroinform.* 8, 80 (2014).

14. Maruyama,R. *et al.* Detecting cells using non-negative matrix factorization on calcium imaging data. *Neural Netw.* 55, 11-19 (2014).

15. Mukamel,E.A., Nimmerjahn,A., & Schnitzer,M.J. Automated analysis of cellular signals from large-scale calcium imaging data. *Neuron* 63, 747-760 (2009).

16. Pacitariu,M. *et al.* Suite2p: beyond 10,000 neurons with standard two-photon microscopy. *BioRxiv*(2017).

17. Pnevmatikakis,E.A. *et al.* Simultaneous Denoising, Deconvolution, and Demixing of Calcium Imaging Data. *Neuron* 89, 285-299 (2016).

18. Reynolds,S. *et al.* ABLE: An Activity-Based Level Set Segmentation Algorithm for Two-Photon Calcium Imaging Data. *eNeuro.* 4, (2017).

19. Spaen,Q. *et al.* HNCcorr: A Novel Combinatorial Approach for Cell Identification in Calcium-Imaging Movies. *eNeuro.* 6, (2019).

20. Stringer,C., Wang,T., Michaelos,M., & Pachitariu,M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods*(2020).

21. Apthorpe,N.J. *et al.* Automatic Neuron Detection in Calcium Imaging Data Using Convolutional Networks. *arXiv*(2016).

22. Soltanian-Zadeh,S., Sahingur,K., Blau,S., Gong,Y., & Farsiu,S. Fast and robust active neuron segmentation in two-photon calcium imaging using spatiotemporal deep learning. *Proc. Natl. Acad. Sci. U. S. A* 116, 8554-8563 (2019).

23. Ronneberger,O., Fischer,P., & Brox,T. U-net:Convolution networks for Biomedical Image Segmentation. *arXiv*(2015).

24. Ohki,K., Chung,S., Ch'ng,Y.H., Kara,P., & Reid,R.C. Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex. *Nature* 433, 597-603 (2005).

25. Smith,S.L. & Hausser,M. Parallel processing of visual space by neighboring neurons in mouse visual cortex. *Nat. Neurosci.* 13, 1144-1149 (2010).

26. Grosenick,L., Marshel,J.H., & Deisseroth,K. Closed-loop and activity-guided optogenetic control. *Neuron* 86, 106-139 (2015).

27. Zhang,Z., Russell,L.E., Packer,A.M., Gauld,O.M., & Hausser,M. Closed-loop all-optical interrogation of neural circuits in vivo. *Nat. Methods* 15, 1037-1040 (2018).

28. Dombeck,D.A., Khabbaz,A.N., Collman,F., Adelman,T.L., & Tank,D.W. Imaging large-scale neural activity with cellular resolution in awake, mobile mice. *Neuron* 56, 43-57 (2007).

29. Bovetti,S. & Fellin,T. Optical dissection of brain circuits with patterned illumination through the phase modulation of light. *J. Neurosci. Methods* 241, 66-77 (2015).

30. Panzeri,S., Harvey,C.D., Piasini,E., Latham,P.E., & Fellin,T. Cracking the Neural Code for Sensory Perception by Combining Statistics, Intervention, and Behavior. *Neuron* 93, 491-507 (2017).

31. Carrillo-Reid,L., Yang,W., Kang Miller,J.E., Peterka,D.S., & Yuste,R. Imaging and Optically Manipulating Neuronal Ensembles. *Annu. Rev. Biophys.* 46, 271-293 (2017).

32. Lecoq,J., Orlova,N., & Grewe,B.F. Wide. Fast. Deep: Recent Advances in Multiphoton Microscopy of In Vivo Neuronal Activity. *J. Neurosci.* 39, 9042-9052 (2019).

33. Keemink,S.W. *et al.* FISSA: A neuropil decontamination toolbox for calcium imaging signals. *Sci. Rep.* 8, 3493 (2018).

34. Lin,T.Y., Goyal,P., Girshick,R., He,K., & Dollar,P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern. Anal. Mach. Intell.* 42, 318-327 (2020).

35. Dubbs,A., Guevara,J., & Yuste,R. moco: Fast Motion Correction for Calcium Imaging. *Front Neuroinform.* 10, 6 (2016).

36. Weiss,K., Khoshgoftaar,T.M., & Wang.D.D. A survey of transfer learning. *Journal of Big Data* 3, (2016).

37. Cogswell,M., Ahmed,F., Girshick,R., Zitnick,L., & Batra,D. Reducing overfitting in deep networks by decorrelating representations. *ICLR Conference Track Proceedings*(2016).

38. Charles,A.S., Song,A., Gauthier,J.L., Pillow,J.W., & Tank,D.W. Neuronal Anatomy and Optical Microscopy (NAOMi) Simulation for evaluating calcium imaging methods. *BioRxiv*(2019).

39. Cicek,O., Abdulkadir,A., Lienkamp,S.S., Brox,T., & Ronneberger,O. 3D U-net: learning dense columetric segmentation from sparse annotation. *Proceedings of the*

International Conference on Medical Image Computing and Computer-Assisted Intervention(2016).

40. Goard,M. & Dan,Y. Basal forebrain activation enhances cortical coding of natural scenes. *Nat. Neurosci.* 12, 1444-1449 (2009).

41. Vinck,M., Batista-Brito,R., Knoblich,U., & Cardin,J.A. Arousal and locomotion make distinct contributions to cortical activity patterns and visual encoding. *Neuron* 86, 740-754 (2015).

42. Jacobs,E.A.K., Steinmetz,N.A., Peters,A.J., Carandini,M., & Harris,K.D. Cortical State Fluctuations during Sensory Decision Making. *Curr. Biol.*(2020).

43. Grewe,B.F. *et al.* Neural ensemble dynamics underlying a long-term associative memory. *Nature* 543, 670-675 (2017).

44. Shuman,T. *et al.* Breakdown of spatial coding and interneuron synchronization in epileptic mice. *Nat. Neurosci.* 23, 229-238 (2020).

45. Villette,V. *et al.* Ultrafast Two-Photon Imaging of a High-Gain Voltage Indicator in Awake Behaving Mice. *Cell* 179, 1590-1608 (2019).

46. Weisenburger,S. *et al.* Volumetric Ca(2+) Imaging in the Mouse Brain Using Hybrid Multiplexed Sculpted Light Microscopy. *Cell* 177, 1050-1066 (2019).

47. Pluta,S.R., Telian,G.I., Naka,A., & Adesnik,H. Superficial Layers Suppress the Deep Layers to Fine-tune Cortical Coding. *J. Neurosci.* 39, 2052-2064 (2019).

48. Antonini,A. *et al.* Extended field-of-view ultrathin microendoscopes for high-resolution two-photon imaging with minimal invasiveness. *Elife.* 9, (2020).

49. Brondi,M. *et al.* High-Accuracy Detection of Neuronal Ensemble Activity in Two-Photon Functional Microscopy Using Smart Line Scanning. *Cell Rep.* 30, 2567-2580 (2020).

50. Forli,A. *et al.* Two-Photon Bidirectional Control and Imaging of Neuronal Excitability with High Spatial Resolution In Vivo. *Cell Rep.* 22, 3087-3098 (2018).

51. Vecchia,D. *et al.* Temporal Sharpening of Sensory Responses by Layer V in the Mouse Primary Somatosensory Cortex. *Curr. Biol.*(2020).

52. Mori,T. *et al.* Inducible gene deletion in astroglia and radial glia--a valuable tool for functional and lineage analysis. *Glia* 54, 21-34 (2006).

53. Castello-Waldow,T.P. *et al.* Hippocampal neurons with stable excitatory connectivity become part of neuronal representations. *PLoS. Biol.* 18, e3000928 (2020).

54. Dombeck,D.A., Harvey,C.D., Tian,L., Looger,L.L., & Tank,D.W. Functional imaging of hippocampal place cells at cellular resolution during virtual navigation. *Nat. Neurosci.* 13, 1433-1440 (2010).

55. Moretti,C., Antonini,A., Bovetti,S., Liberale,C., & Fellin,T. Scanless functional imaging of hippocampal networks using patterned two-photon illumination through GRIN lenses. *Biomed. Opt. Express* 7, 3958-3967 (2016).

56. He,K., Zhang,X., Ren,S., & Sun,J. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*(2016).

57. Kingma,D.P. & Ba,J.L. Adam: a method for stochastic optimization. *arXiv*(2015).

58. Herlihy,M. & Shavit,N. *The art of multiprocessor programming*(Elsevier,2006).

59. Malisiewicz,T., Gupta,A., & Efros,A.A. Ensemble of exemplar-SVMs for object detection and beyond. *Proceedings of the IEEE International Conference on Computer Vision*(2011).

60. Abadi,M. *et al.* TensorFlow: a system for large-scale machine learning. *arXiv*(2016).

61. Van Rossum,G. & Drake,F.L. *Python 3 Reference Manual*(CreateSpace,2009).

62. Miller,J.E., Ayzenshtat,I., Carrillo-Reid,L., & Yuste,R. Visual stimuli recruit intrinsically generated cortical ensembles. *Proc. Natl. Acad. Sci. U. S. A* 111, E4053-E4061 (2014).

Methods

Animals

All experiments were carried out in accordance with the guidelines of the European Communities Council Directive, and were approved by the National Council on Animal Care of the Italian Ministry of Health (authorization #34/2015-PR, #1134/2015-PR, and #61/2019-PR).

Wild type (wt) C57BL/6J mice were purchased from Charles River Laboratories (Calco, Italy; strain code #632), transgenic *Scnn1a-cre* (B6;C3-Tg(Scnn1a-cre)3Aibs/J; JAX #009613), and *Ai95D* (B6;129S-*Gt(ROSA)26Sor*$^{tm95.1(CAG-GCaMP6f)Hze}$/J; JAX #024105) were purchased from Jackson Laboratories (Bar Harbor, USA). *Scnn1a-cre* mice express the enzyme Cre in a subpopulation of layer IV neurons [47] and of VPM cells [48]. Animals were housed in individually ventilated cages under a 12-hr light:dark cycle. A maximum of five animals per cage was allowed with *ad libitum* access to food and water. Mice of both sexes were used for experiments.

Viral injections and surgical procedures

We expressed GCaMP6 or GCaMP7 through the following viral vectors AAV1.Syn.Flex.GCaMP6s.WPRE.SV40 (Addgene viral prep # 100845-AAV1), AAV1.Syn.Flex.GCaMP6f.WPRE.SV40 (Addgene viral prep # 100833-AAV1) purchased from the University of Pennsylvania Viral Vector Core, or AAV1.Syn.Flex.GCaMP7f.WPRE.SV40 (Addgene viral prep #104492-AAV1) purchased from Addgene. For CA1 imaging, we expressed jRCaMP1a using co-injection of AAV1.CAG.Flex.NES-jRCaMP1a.WPRE.SV40 (Addgene viral prep # 100846-AAV1) and AAV1.CamKII 0.4.Cre.SV40 (Addgene viral prep # 105558-AAV1) purchased from the University of Pennsylvania Viral Vector Core.

For LIV imaging, we used a total of 29 mice. Specifically, 17 *Scnn1a-cre* mice injected with a viral vector transducing GCaMP6s, 6 *Scnn1a-cre* mice injected with a virus transducing GCaMP6f, 3 *Scnn1a-cre* mice injected with a virus transducing GCaMP7s, and 3 *Scnn1a-cre* crossed with *Ai95D* mice. Mice between post-natal days 30 and 33 were anesthetized with 2% isoflurane (IsoFlu, Zoetis, IT) in 0.8 % oxygen, placed into a stereotaxic apparatus (Stoelting Co, Wood Dale, IL), and maintained on a warm platform at 37°C for the whole duration of the anesthesia. Viral injection in mice used for LIV imaging was carried out similarly to [49] and [50]. Briefly, a scalp incision was performed after local administration of lidocaine (2 %) and then two small holes were drilled on the skull above the right/left somatosensory cortex at 1.2 mm and 2 mm posterior (P) to the bregma suture, 2.8 mm and 3 mm lateral (L) to the sagittal sinus. A micropipette was slowly inserted into the cortical tissue until the tip reached a depth of 0.3 mm below the pia mater. 200 nL of GCaMP6 virus were injected at 20 - 60 nl/min by means of a hydraulic injection apparatus driven by a syringe pump (UltraMicroPump, WPI, Sarasota, FL). The pipette was then further lowered to reach 0.4 mm below the *pia mater*, and a second injection was performed. This procedure was repeated for the second injection site. The injected solution contained 10^{12} viral genomes/ml diluted 1:1 in artificial cerebro-spinal fluid (aCSF: 127 mM NaCl, 3.2 mM KCl, 2 mM $CaCl_2$, 1 mM $MgCl_2$ and 10 mM HEPES, pH 7.4). The exposed skull was then cleaned, and the skin sutured and cleansed with Iodopovidone (Betadine®, Meda pharma, Milan, Italy). Mice were monitored until full recovery from the anesthesia. In mice used for imaging in awake conditions, a custom metal bar was sealed to the skull using Vetbond (3 M, St. Paul, MN, USA) and dental cement (Paladur, Kulzer GmbH, Hanau, Germany). The exposed bone was covered using the silicone elastomer KWIK-Cast (World Precision Instruments, Friedberg, DE) and an intraperitoneal injection of antibiotic (BAYTRIL, Bayer, DE) was performed. 2-4 weeks after virus injection, mice used for imaging in LIV during anesthesia were injected with urethane (1.65 g/kg, 16.5 % in saline solution, i-p.). A scalp incision was performed

after local administration of Lidocaine (2 %). A circular craniotomy was opened over the somatosensory cortex, in the area where green fluorescence was clearly visible using an epifluorescent microscope. The surface of the brain was kept moist with aCSF. A heating pad underneath the animal was set at 35.5-37°C. Respiration rate, eyelid reflex, vibrissae movements, and reactions to tail pinching were monitored throughout the surgery. Mice were then moved under the two-photon microscope, kept at 37 °C with a heating pad, and the brain surface irrigated with aCSF. Imaging began one hour after the end of the surgery. Before imaging LIV activity in awake animals, mice were habituated to head-fixation similarly to [51]. In brief, habituation lasted for 7 to 10 days, during which they were head restrained for increasing periods (from 15 minutes to one hour), while running or standing on a custom made treadmill. On the day of the experiment, the habituated mouse was anesthetized with a mixture of isoflurane and oxygen (0.8 – 2 %), and a craniotomy performed similarly to what described above. After surgery, the animal was head fixed and allowed to recover under the microscope for at least one hour before imaging.

For VPM imaging, we used a total of 4 mice. Viral injections and aberration-corrected microendoscopes insertion in mice used for VPM imaging were performed in Scnn1a-Cre mice as in [48]. Mice were anesthetized as previously described. A single craniotomy was performed at stereotaxic coordinates P 1.7 mm, L 1.6 mm. A micropipette was lowered to a depth of 3 mm below the brain surface. 0.5 - 1 µl of GCaMP6s virus containing solution (containing 10^{12} viral genomes/ml diluted 1:4 in aCSF) were injected at 30 - 50 nl/min. Following virus injection, a craniotomy (area: 600 µm x 600 µm) was performed at stereotaxic coordinates P 2.3 mm, L 2 mm. A thin column of brain tissue was displaced with a glass cannula (ID = 300 µm, OD = 500 µm; Vitrotubs, Vitrocom Inc., Mounting Lakes, NJ) and a microendoscope was slowly inserted into the cannula track using a custom holder,

down to 3 mm from the brain surface. The microendoscope was finally secured by acrylic adhesive and dental cement. Imaging was performed 2-4 weeks after endoscope implantation.

For CA1 imaging, we used a total of 2 mice. Before surgery, mice were medicated with an intramuscular bolus of Dexamethasone (Dexadreson, 4 gr/kg). After scalp incision, a 0.5 mm craniotomy was drilled at stereotaxic coordinates P 1.75 mm, L 1.35 mm. A micropipette was lowered 1.40 mm below the brain surface. 0.8 µl of viral solution (containing a mixture of CamKII-Cre and jRCaMP1a viruses at 10^{12} viral genomes/ml diluted 1:1 in aCSF) was injected at 100 nL/min in *Ai95* crossed with *Glast-cre-ERT2* mice [52]. Inducible Glast-cre was not activated after viral injection, resulting in CA1 neuronal staining with jRCaMP1a and GCaMP6f. A stainless-steel screw was attached to the skull, and a chronic hippocampal window was implanted as described in [53, 54]. A 3 mm craniotomy was opened, centered at coordinates P 2.00 mm, L1.80 mm. The *dura mater* was removed using fine forceps, and the cortical tissue overlaying the hippocampus slowly aspirated using a blunt needle coupled to a vacuum pump. During aspiration, the exposed tissue was continuously irrigated with aCSF. Aspiration was stopped once the fibers of the external capsule were exposed. A cylindrical optical window made of a thin-walled stainless-steel cannula (OD, 3 mm; ID, 2.77 mm; height, 1.50 - 1.60 mm) attached to a 3 mm diameter round coverslip, was fitted to the craniotomy in contact to the external capsule. A thin layer of silicone elastomer was used to surround the interface between the brain tissue and the steel surface of the optical window. A custom stainless-steel headplate was attached to the skull using epoxy glue. All the components were finally fixed in place using black dental cement and the scalp incision was sutured to adhere to the implant. All the animals received an intraperitoneal bolus of antibiotic (BAYTRIL, Bayer, DE) to prevent postsurgical infections.

762

763 *Functional two-photon imaging*

764 Two-photon imaging was performed using a chameleon ultra II pulsed laser (80 MHz pulse frequency,
765 Coherent Inc, Santa Clara, CA, USA) tuned at 920 nm for GCaMP6/7 imaging and at 990 nm for
766 dual color imaging. Excitation power was 30 - 110 mW as measured under the microscope objective
767 and controlled via a Pockel cell (Conoptics Inc, Danbury CT, USA,). An Ultima II scanhead (Bruker
768 Corporation, Milan, Italy) equipped with 3 mm raster scanning galvanometers (6215H, Cambridge
769 Technology, Bedford, MA) and standard photomultiplier tubes (Hamamatsu Photonics, Tokyo, Japan)
770 and an Ultima Investigator (Bruker Corporation, Milan, Italy), equipped with 6 mm raster scanning
771 galvanometers, movable objective mount, and multi-alkali photomultiplier tubes were used. The three
772 objectives were: 25x/1.05 N.A. (Olympus Corp., Tokyo, JP) for LIV imaging, 20x/0.5NA (Zeiss,
773 Oberkochen, Germany) for VPM endoscopic imaging, and 16x/0.8 N.A. (Nikon, Tokyo, Japan) for
774 CA1 experiments.

775

776 For LIV imaging, dwell time was 4.4 µs, photomultiplier voltage was 777 V, zoom factor was always
777 1, pixel size was 0.77 µm, acquisition frame rate ranged between 0.5 – 3 Hz for a 512 pixels x 512
778 pixels image. Fluorescence values spanned 95 % of the available dynamic range (16 bit). For dual
779 color CA1 imaging, pixel dwell-time was set at 4 µs, photomultiplier voltage was 777 V, zoom factor
780 was always 5, pixel size was 0.634 µm, acquisition frame rate was 3.03 Hz for a 256 pixels x 256
781 pixels image. For VPM imaging, the set up was similar to the one described in [48, 55], pixel dwell-time
782 was set at 4 µs, photomultiplier voltage was 810 V, zoom factor was always 1, pixel size was 2.19
783 µm, acquisition frame rate was 2.66 Hz for a 196 pixels x 196 pixels image. Imaging sessions lasted
784 1 hour for CA1, VPM, and awake LIV experiments. They lasted 4 hours for the anesthetized LIV
785 condition. After awake imaging sessions, animals were returned to their home cages.

Training and validation datasets

In the absence of a generally accepted wide-scale annotated dataset of two-photon calcium imaging, we built a dataset of *in vivo* t-series collected using raster scanning acquisitions. A total of 197 t-series (average frame number per time series: 597 ± 262, average frame rate: 2.3 ± 1.5 Hz) were included in the dataset: 76 t-series from CA1 imaging of principal neurons stained with both GCaMP6f (38 t-series) and jRCaMP1a (38 t-series); 121 t-series from cortical LIV imaging of principal neurons stained with virally injected GCaMP6s (107 t-series), GCaMP6f (4 t-series), GCaMP7f (5 t-series), GCaMP6f expressed in transgenic animals (Scnn1a-cre x Ai95; 5 t-series). Training and validation datasets contained 160 (118 from LIV, 21 from CA1 GCaMP6f, and 21 from CA1 jRCaMP1a) and 37 t-series (13 from LIV, 12 from CA1 GCaMP6f, and 12 from CA1 jRCaMP1a), respectively. To avoid data leakage between training and validation datasets, we manually split t-series including different FOVs in the datasets.

Additional datasets

Four additional datasets were selected and used for validation purposes only:

1) VPM microendoscopic imaging t-series in awake head restrained mice (9 t-series).

2) The publicly available Allen Brain Observatory (ABO) visual coding dataset (19 t-series, https://observatory.brain-map.org/visualcoding). T-series identification numbers: 501271265, 501484643, 501574836, 501704220, 501729039, 501836392, 502115959, 502205092, 502608215, 503109347, 504637623, 510214538, 510514474, 510517131, 527048992, 531006860, 539670003, 540684467, 545446482.

3) The publicly available Neurofinder (NF) challenge dataset (28 t-series, https://github.com/codeneuro/neurofinder). T-series identification numbers: neurofinder.00.00,

neurofinder.00.01, neurofinder.00.02, neurofinder.00.03, neurofinder.00.04, neurofinder.00.05, neurofinder.00.06, neurofinder.00.07, neurofinder.00.08, neurofinder.00.09, neurofinder.00.10, neurofinder.00.11, neurofinder.01.00, neurofinder.01.01, neurofinder.02.00, neurofinder.02.01, neurofinder.03.00, neurofinder.04.00, neurofinder.04.01, neurofinder.00.00.test, neurofinder.00.01.test, neurofinder.01.00.test, neurofinder.01.01.test, neurofinder.02.00.test, neurofinder.02.01.test, neurofinder.03.00.test, neurofinder.04.00.test, neurofinder.04.01.test.

4) A single t-series of mesoscopic full field imaging from [7].

No preprocessing was performed on the VPM, ABO, NF, and mesoscopic t-series. All t-series were manually annotated *de novo* by the two graders working independently. The consensus ground truth was obtained as described for the training and validation dataset (see also Supplementary Table 1-2).

Image processing and consensus labelling

The CITE-On image detector was based on purely morphological features extracted from imaging data. No information from the dynamic fluorescence signal in the t-series was used to detect putative cells. Each imaging t-series was corrected for lateral displacements using MOCO [35]. T-series were aligned using the first frame as reference, without downsampling and with a maximum possible shift of 13 pixels. The 8-bit median projection of each t-series was then computed on the motion-corrected t-series. The resulting images (one *per* t-series) were globally sharpened to better visualize cell shapes. A gamma correction of 0.3 was applied, and the dynamic range was linearly adjusted normalizing across the whole 8 bit depth. Processed images were named “enhanced median projections” (EMPs) and were used to define our GT labelling. Two graders independently labelled each EMP. LabelImg (http://github.com/tzutalin/labelImg) was used to define a single object class by manually drawing bounding boxes around every visible cell soma in the EMP. The surface of each bounding box was manually defined in order to tightly surround the cell shape. Boxes were allowed

to overlap. Coordinates and surface of each bounding box for all EMPs were saved in a standard VOC format where each file reported the top left and bottom right coordinates (in pixels) for each bounding box. The GT was defined as the union between the set of notations from the two graders.

Image detector training

CITE-On is based on a fully convoluted single-shot image detector, RetinaNet [34]. Briefly, a feature pyramid network was constructed from residual layers of the ResNet50 feature extractor [56]. This feature pyramid was then fed to two separate sets of convolution filters: one computing the label score (classification subnet), the second performing bounding box regression from priors (regression subnet). We used the open-source Keras implementation of this architecture (http://github.com/fizyr/keras-retinanet). Given the relatively small set of training data available, we could not perform *ab initio* training. Rather, we employed a transfer learning approach. Starting from the network trained on a large-scale dataset of natural images, we fine-tuned the weights of the classification and regression subnets, while freezing the weights of the feature pyramids. We used "plain" median projections obtained from the motion corrected t-series and linearly normalized across the bit range. The resulting projections were then upsampled in order to obtain images were the short side was 800 pixels long, while the long side did not exceed 1,333 pixels. Since the input layer of the network accepted a three-channel image, the same image was replicated for each channel without changing any parameter. These last two image conversions were necessary as the network was originally trained on RGB images, and the transfer learning approach did not allow input modifications to the image shape. The network was trained with a regression L1 loss function (Mean Absolute Errors, MAE, https://rishy.github.io/ml/2015/07/28/l1-vs-l2-loss/), with focal loss (http://arxiv.org/abs/1708.02002) and the Adam optimizer [57] with learning rate 10-5 and clipnorm 10-3 (http://github.com/keras-team/keras/issues/510) [34] modified by reducing the learning rate on loss

plateau with a factor of 0.1. The network was trained for 17 epochs, each consisting of 1000 training steps of batch size 1.

CITE-On offline pipeline

Two-photon calcium imaging t-series were first corrected for lateral artifacts using MOCO (as described above). The median projections were then computed, normalized, upscaled to the target input size, and converted to 8 bit RGB images. The resulting images were fed to the image detection network. Upscaling factor and score threshold were the only two parameters defined *a priori*. Upscaling factor was adjusted in order to have the smallest feature in each image inscribed in a 32 $pixels^2$ box. This was because the smallest prior box encoded in the network was 32 $pixels^2$. To avoid using a shallower convolutional residue that would have carried less information, we decided to adjust the up-scaling factor instead of reducing the size of the smallest prior box. After exploring the parameter space on different datasets, we defined an up-scaling factor of 1 for the training and validation datasets. The optimal up-scaling factors were 2.3 for the ABO dataset, and between 1.7 and 3.1 for the NF datasets (these datasets were heterogeneous in terms of FOV size and pixel size). In the VPM dataset, image projections were characterized by an altered magnification factor in the radial direction, due to the optical properties of the corrected microendoscopes used for imaging [48]. These images went through a further preprocessing step to correct for this field distortion. Up-scaling factor for the VPM dataset was 1.4. Each bounding box was associated with a score, representing network confidence in cell detection. Bounding boxes with intersection over union (IoU) < 20 % were considered as separate neuronal identities. When IoU of two bounding boxes was > 20 %, the bounding box with the highest score was retained. Results of the image detector were filtered by applying a threshold on the output score provided by the network and optimized for each dataset.

CITE-On online pipeline

In the online pipeline, individual raw imaging frames were continuously grabbed from a streaming source (e.g. live microscope output) and processed. To simulate this process, we individually imported in the CITE-On pipeline each frame of each raw t-series. Single frames were passed on to the trace extractor and to a buffer. The buffer stored the number of frames sufficient to produce an average projection. Once the buffer was filled, the projection was computed, sent to the image detector, and the buffer emptied.

Detections were performed using the same procedure described for the offline pipeline. Detection results were fed to a custom tracking algorithm detecting all the overlaps between current and previous detections, and designed in order to maximize overlap between putative matching boxes. For every detection matching a previous one, the coordinates of the relative bounding box were updated to the last one. For each new detection having an IoU < 0.25 % with all the previous detections, a new identity was created. In case of identities not actively detected in the current frame, relative coordinates were anyway updated using a rigid shift calculated as the mean shift obtained from the active identities. In this way, we aimed to minimize the effect of motion artifacts and identity switch without implementing online motion correction approaches. A simple dynamic segmentation was then performed for each identity. At each raw frame, the interval between the 80^{th} and the 95^{th} percentile of the pixel fluorescence intensity distribution inside each bounding box was averaged to extract the raw functional trace. At each frame, background signal corresponded to the average fluorescence of all the pixels in the FOV not belonging to any bounding box. This frame-wise value was subtracted from all the individual raw functional traces. In order to optimize real-time performance for high frame rate acquisitions (above 3 Hz, including all ABO and some NF t-series), the entire pipeline was implemented in a multiprocessing scheme where one process was responsible

for data loading, one for image preprocessing, and for sending its output to the CNN detector accelerated over GPU and tracking parts, while the remaining CPU cores were dedicated to real-time trace extraction given the parallel nature of the problem [58]. This implementation scheme allowed for cell detection update (up to 10 Hz) and functional trace extraction update from all identities (100 Hz) to operate as parallel and asynchronous processes.

Parameter exploration for the object detector

To find the best operating parameters for the object detector, we quantified offline CITE-On performance while systematically exploring various plausible values of score threshold and upscaling factor. For all datasets, we evaluated the F-1 across a set of net size multipliers between 0.6 and 3.4 in steps of 0.2. We also explored score thresholds between 0.05 and 0.95, in steps of 0.05. This mapping strategy allowed us to define the optimal combination of score threshold and upscaling factor for each input data. The upscaling factor was dependent on the ratio between the average box surface and the whole FOV surface, while the score threshold presented non obvious dependence on trivial image statistics. Therefore, we determined the upscaling factor according to the acquisition parameters and the relative score thresholds, in order to maximize the F-1 score for each dataset. For the online pipeline, we used appropriate upscaling factors and proceeded by exploring the dependency of F-1 from the score threshold and from the number of averaged frames in each detection.

Trace extraction: comparison between CITE-On and other methods

We compared CITE-On trace extraction with trace extraction performed with a popular state-of-the-art method based on CNMF, CaImAn [10]. Briefly, we provided binary masks corresponding to the CITE-On detected bounding boxes and used these masks to seed the CNMF algorithm. Seeded-CNMF first calculated the temporal background component using pixels that were not included in

any mask. We compared this background component to the CITE-On background traces used for trace correction. The subsequent step of the seeded-CNMF algorithm estimated temporal components and spatial footprints, constrained to be non-zero only at the location of the binary masks. Using this strategy, we obtained fluorescent traces from putative cells detected in the same locations as those detected by CITE-On, allowing for a one-to-one trace comparison between algorithms. It is important to note that, for this analysis, we set the order of the autoregressive model of the CNMF to zero because we were not interested in trace deconvolution, but only in correcting for background contamination.

Local vs. global background signal correlation

To compare local and global background noise contributions, we used the same approach for background noise subtraction, but considering only the pixels in the vicinity of each cell. The vicinity of a cell was defined as all the pixels in a concentric rectangular box double the size of the box detected by CITE-On, with no overlap with other bounding boxes. We then calculated the cross correlation at lag zero between the local noise for each cell and the global background trace.

Tiled detection on mesoscopic images

For large scale datasets (> 3000 pixel^2) as the mesoscopic imaging dataset [7] , requiring large amounts of GPU memory (> 500 GB), we implemented a tiled detection approach. We divided each mesoscopic FOV in a configurable number of tiles with configurable overlap factor, in order to batch process all tiles up to the limit of the available GPU memory. Once all detections were computed, they were appropriately shifted back to the original position in the FOV and a Non Maximum Suppression [59] was performed in order to remove duplicate boxes in regions of the FOV were overlap between tiles occurred.

954

955 *Performance and metrics*

956 Performance of the CITE-On object detector was calculated employing standard Precision, Recall
957 and F-1 metrics. In order to determine true or false positives, we computed the IoU between each
958 ground truth and detected boxes. A custom algorithm was used to match the two sets using a cut-off
959 threshold of 0.5.

960

961 *Hardware and software for data analysis*

962 All the data analysis procedures presented in this work were performed on a Dell Precision 7920
963 desktop with an Intel Xeon Silver 4110 @ 2.1 GHz 8 core CPU, 32 GB DDR4-2666 ECC RAM,
964 NVIDIA Quadro RTX5000 GPU, 512 GB NVMe SSD, and 2 TB 7200 rpm HDD.

965

966 All processing steps, including network training and validation, were carried out under
967 Keras/Tensorflow software libraries [60]. Image processing and data analysis were carried out using
968 Python Language Reference 3 [61].

969

970 *Statistics*

971 Values were expressed as mean ± sd, unless otherwise stated. The number of samples (N) and p values
972 are reported in the figure legends or in the text. No statistical methods were used to pre-determine
973 sample size, but sample size was chosen based on previous studies [3, 62]. All recordings with no
974 technical issues were included in the analysis and blinding was not used in this study. Statistical
975 analysis was performed with the scientific Python ecosystem (SciPy, NumPy 1.19) under Python 3,
976 Python Software Foundation, Python Language Reference 3 (available at https://www.python.org). A
977 Kolmogorov-Smirnov test was run on each experimental sample to test for normality. The

significance threshold was always set at 0.05. When comparing two paired populations of non-normally distributed data, a two-sample Kolmogorov-Smirnov test. All tests were two-sided, unless otherwise stated.

Code availability

The code is available from the corresponding authors on request and it will be shared upon publication.

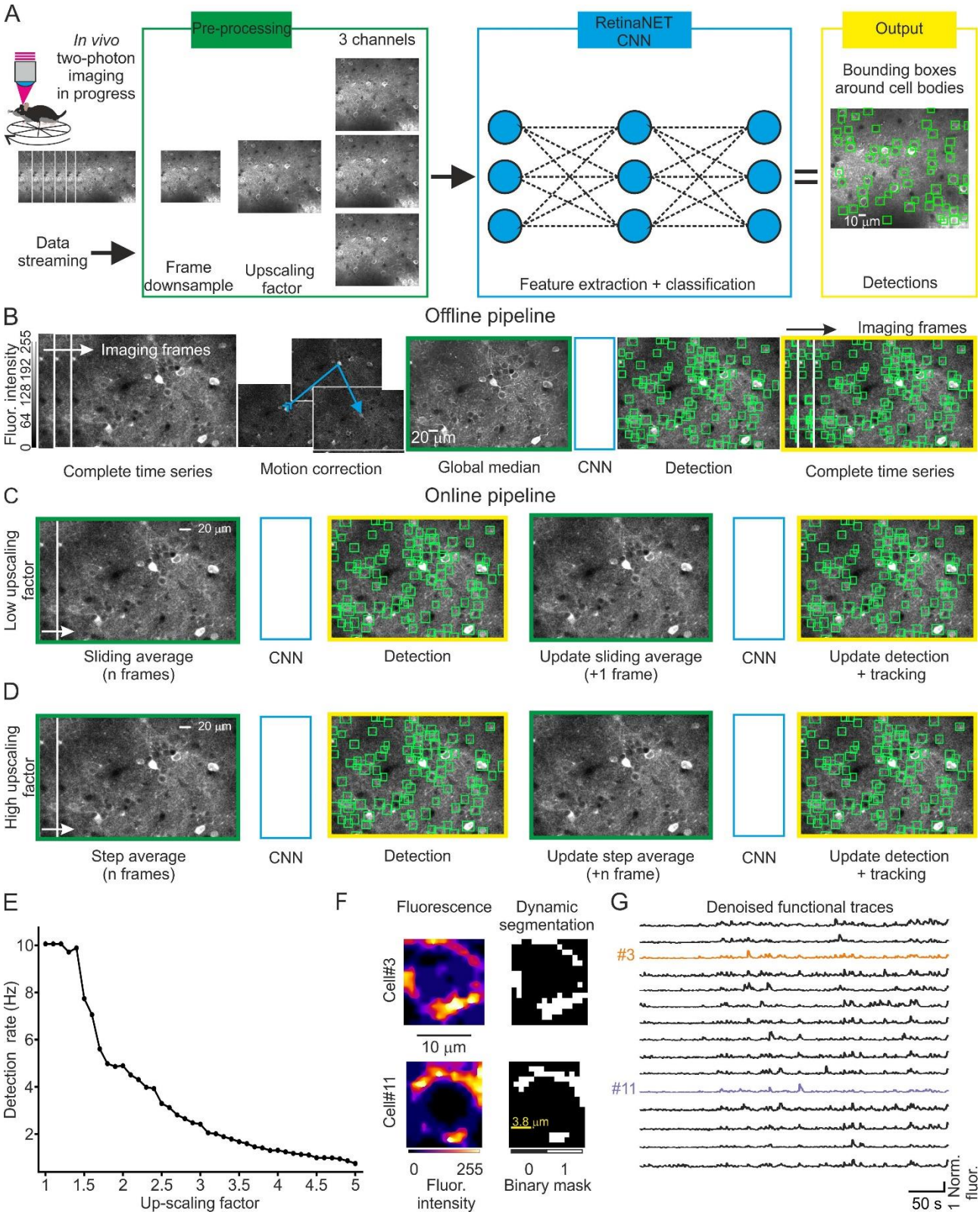


Figure 1. Structure and analysis pipeline of CITE-On. A) Schematic of the image detection process in CITE-On. During ongoing two-photon imaging acquisition, individual frames are

transferred to CITE-On as they are completed (left). A pre-processing step (green rectangle) is required ahead of image detection, comprising frame downsampling, image upscaling and triplication of the upscaled image. The result of the preprocessing is then used as input for the CNN (blue rectangle). The CNN output is the detection of neuronal somata in the form of bounding boxes (green squares black and white image on the right). **B)** CITE-On offline pipeline starts with the complete t-series and the correction of motion artifacts (blue arrow, motion correction). Frame downsampling is performed by computing the global median projection of the t-series. The upscaled and triplicated global median (green) is fed to the CNN (blue), a single detection is performed, and the bounding boxes (detection, green squares) are projected onto each frame of the complete t-series (yellow). **C)** In the online pipeline, for data requiring small upscaling factors, a sliding average projection of the first n frames of the ongoing t-series is calculated in the frame downsampling pre-processing step (green). This image is upscaled and triplicated, processed by the CNN (blue), producing the first detection (yellow). As the next frame of the t-series is acquired, a new sliding average is computed, again on *n* frames, but starting from the second frame of the acquisition and including the $n+1^{th}$ one. The CNN processes this image, updating the detections and starting the tracking system (yellow). The grayscale shown in this panel applies to all grayscale images in this figure. **D)** For data requiring high upscaling factors, the pipeline is similar to that in (C), but instead of a sliding average, a step average is calculated on *n* frames as the frame downsampling pre-processing step (green). Detections are updated every *n* new frames. **E)** Detection rates as a function of the magnitude of the upscaling factor. Maximum detection rate is 10 Hz for upscaling factor between 1 and 1.5. **F)** Representative average fluorescence of pixels inside the bounding box relative to two cells (cell #3 and cell #11), calculated in a single frame of the LIV dataset (GCaMP6s, pseudocolor, left). Associated dynamic segmentation mask in the same frame (binary mask, right). **G)** Functional traces from N = 15

representative cells extracted with online CITE-On pipeline. Traces of cells displayed in (F) are shown in green.

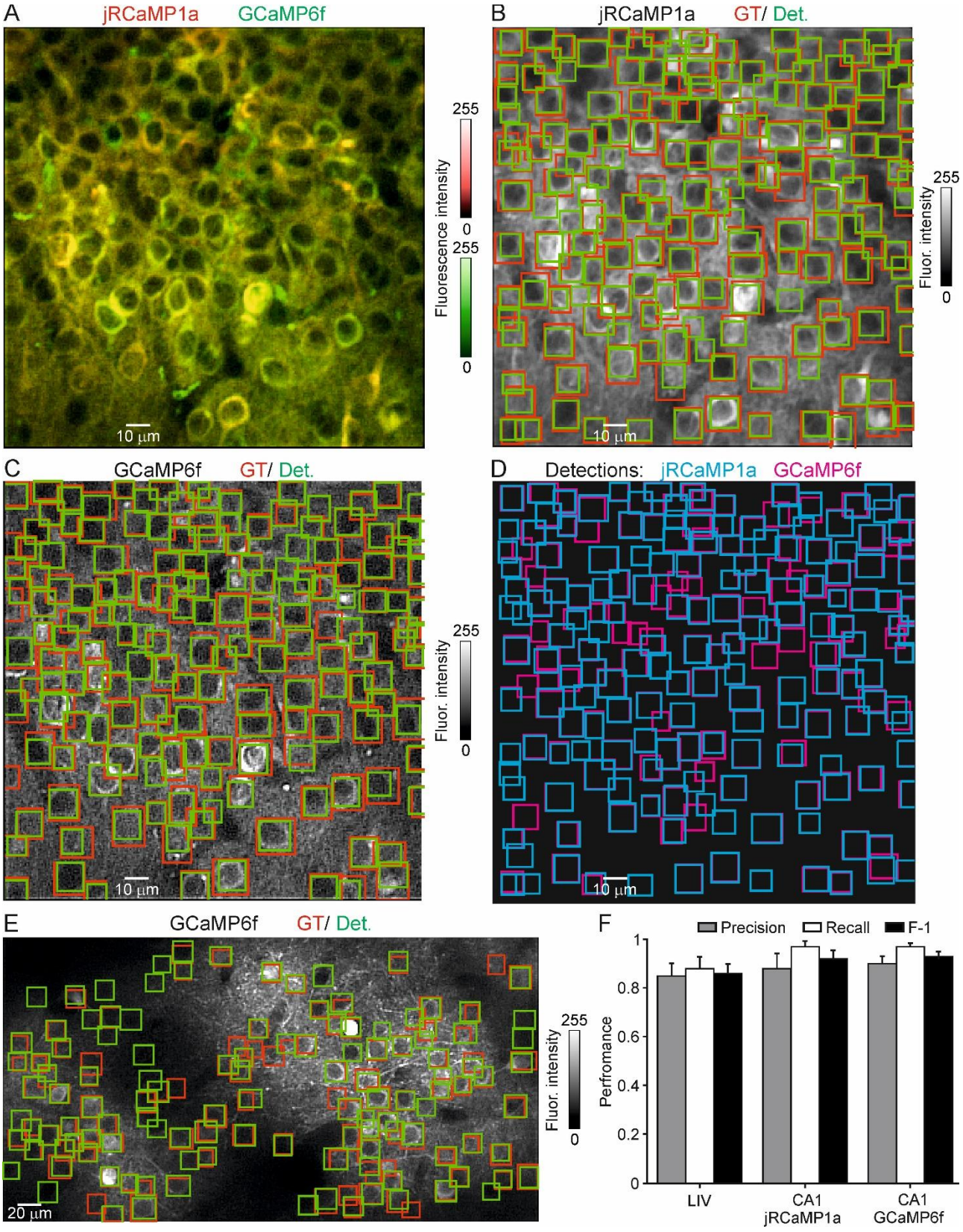


Figure 2. CITE-On offline performance. A) Representative median projection showing jRCaMP1a (red) and GCaMP6f (green) expressing CA1 neurons. **B)** GT (red) and CITE-On detections (green) for the jRCaMP1a channel of the image shown in A. **C)** same as in (B), but for the GCaMP6f channel. **D)** Superposition of CITE-On detections on jRCaMP1a (cyan) and GCaMP6f (magenta) channels. **E)** Representative median projection from the LIV dataset with GT (red) and CITE-On detections (green). **F)** Performance for precision (grey), recall (white), and F-1 (black) obtained with the offline CITE-On pipeline on the validation t-series of the LIV (N = 13), CA1 jRCaMP1a (N = 12), and CA1 GCaMP6f (N = 12) datasets. In this and in other figures values are expressed as means ± sd.

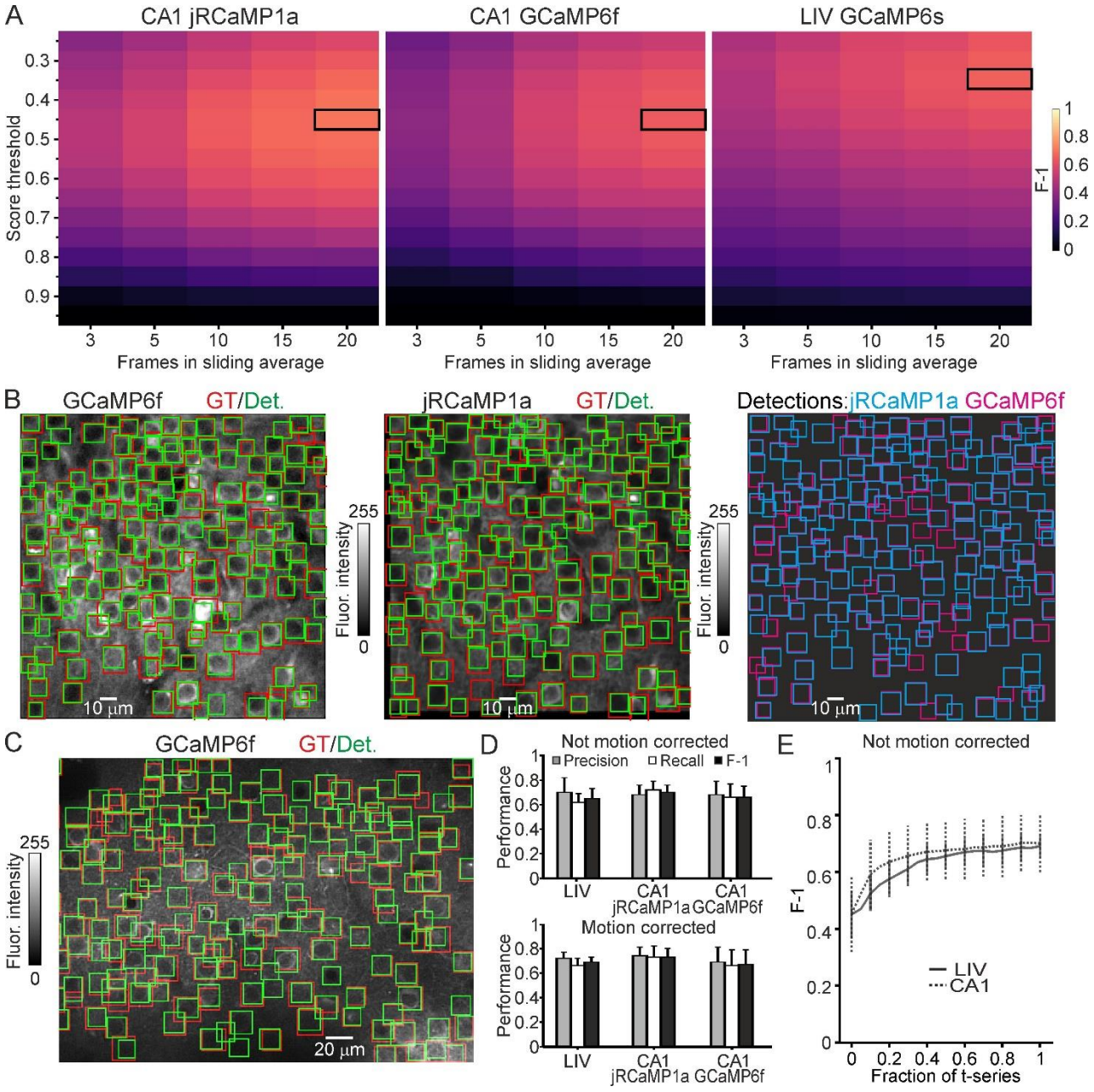


Figure 3. CITE-On online performance. A) Best parameter search for frame downsampling: F-1 score (pseudocolor) as a function of score threshold (vertical axis) and number of frames in the sliding average (horizontal axis) for LIV (left), CA1 GCaMP6f (middle) and CA1 jRCaMP1a (right). The maximal F-1 is indicated with the black rectangle. **B)** Median projections of one representative FOV for CA1 jRCaMP1a (left) and one FOV for CA1 GCaMP6f (middle). GT (red) and online detections (DET, green) are also shown. In the rightmost panel, the online detections of jRCaMP1a (cyan) and GCaMP6f (magenta) are shown. **C)** Same as in (B) but for a representative LIV t-series. **D)** Top:

online performance of Precision (grey), Recall (white), and F-1 (black) for all t-series in the validation LIV (N = 13), CA1 jRCaMP1a (N = 12), and CA1 GCaMP6f (N = 12) datasets. No motion correction was performed. Bottom: same as top, but for the motion corrected t-series. Results of Kolmogorov-Smirnov test for performance in not motion corrected t-series vs. motion corrected t-series from LIV: $p = 0.54$ for F-1, $p = 0.15$ for Precision, $p = 0.38$ for Recall, N = 13 t-series. Results of Kolmogorov-Smirnov test for performance in not motion corrected t-series vs. motion corrected t-series from CA1 jRCaMP1a: $p = 0.16$ for F-1, $p = 0.20$ for Precision, $p = 0.20$ for Recall, N = 12 t-series. Results of Kolmogorov-Smirnov test for performance in not motion corrected vs. motion corrected t-series from CA1 GCaMP6f: $p = 0.18$ for F-1, $p = 0.28$ for Precision, $p = 0.22$ for Recall, N = 12 t-series. **E)** F-1 values as a function of the fraction of the total length of the t-series for not-motion corrected data (N = 13 t-series for LIV, N = 24 t-series for CA1, including N = 12 t-series for CA1 jRCaMP1a and N = 12 t-series for CA1 GCaMP6f t-series).

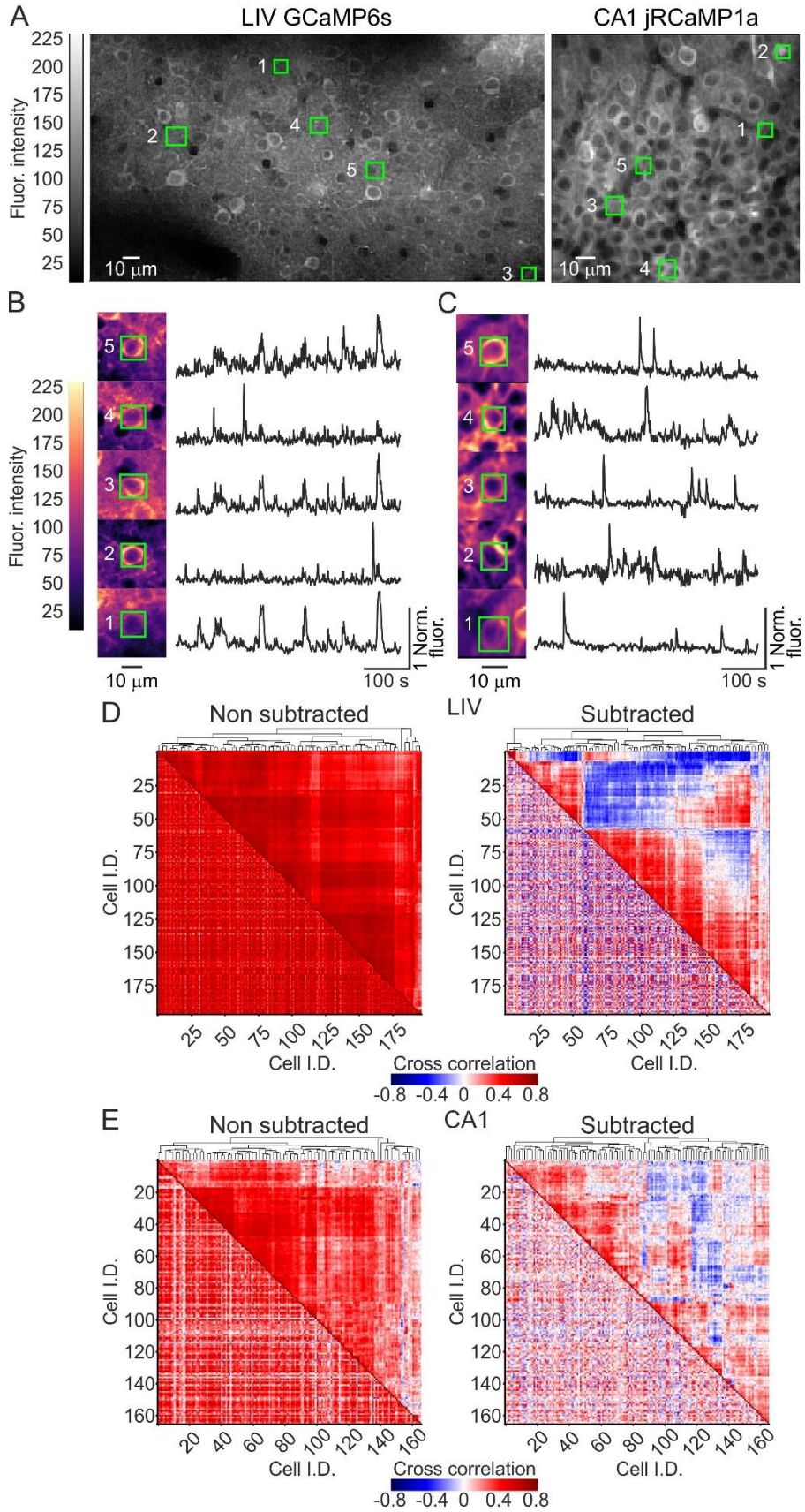

A
LIV GCaMP6s
CA1 jRCaMP1a
Fluor. intensity
225
200
175
150
125
100
75
50
25
10 μm
B
C
Norm. fluor.
1
100 s
D
Non subtracted
Subtracted
LIV
Cell I.D.
Cross correlation
-0.8 -0.4 0 0.4 0.8
E
CA1

Figure 4. Fast extraction of fluorescence traces using CITE-On. A) Median projection showing representative FOVs from the LIV GCaMP6s (left) and the CA1 jRCaMP1a (right) datasets. True positive bounding boxes for five CITE-On identified cells in each FOV are shown. **B)** Left: the five cells indicated in the LIV t-series displayed in (A) are shown at an expanded spatial scale. Right: corresponding background subtracted fluorescence traces. **C)** Same as in B but for the CA1 t-series in A. **D)** Lower-left triangle: cross correlation matrix for all functional traces extracted from true positive detection in the LIV GCaMP6s t-series displayed in (A). Upper-right triangle: corresponding dendrogram sorting. The left matrix shown signals before background subtraction. The right matrix after background subtraction. **E)** Same as in (D), but for the CA1 jRCaMP1a t-series shown in (A).

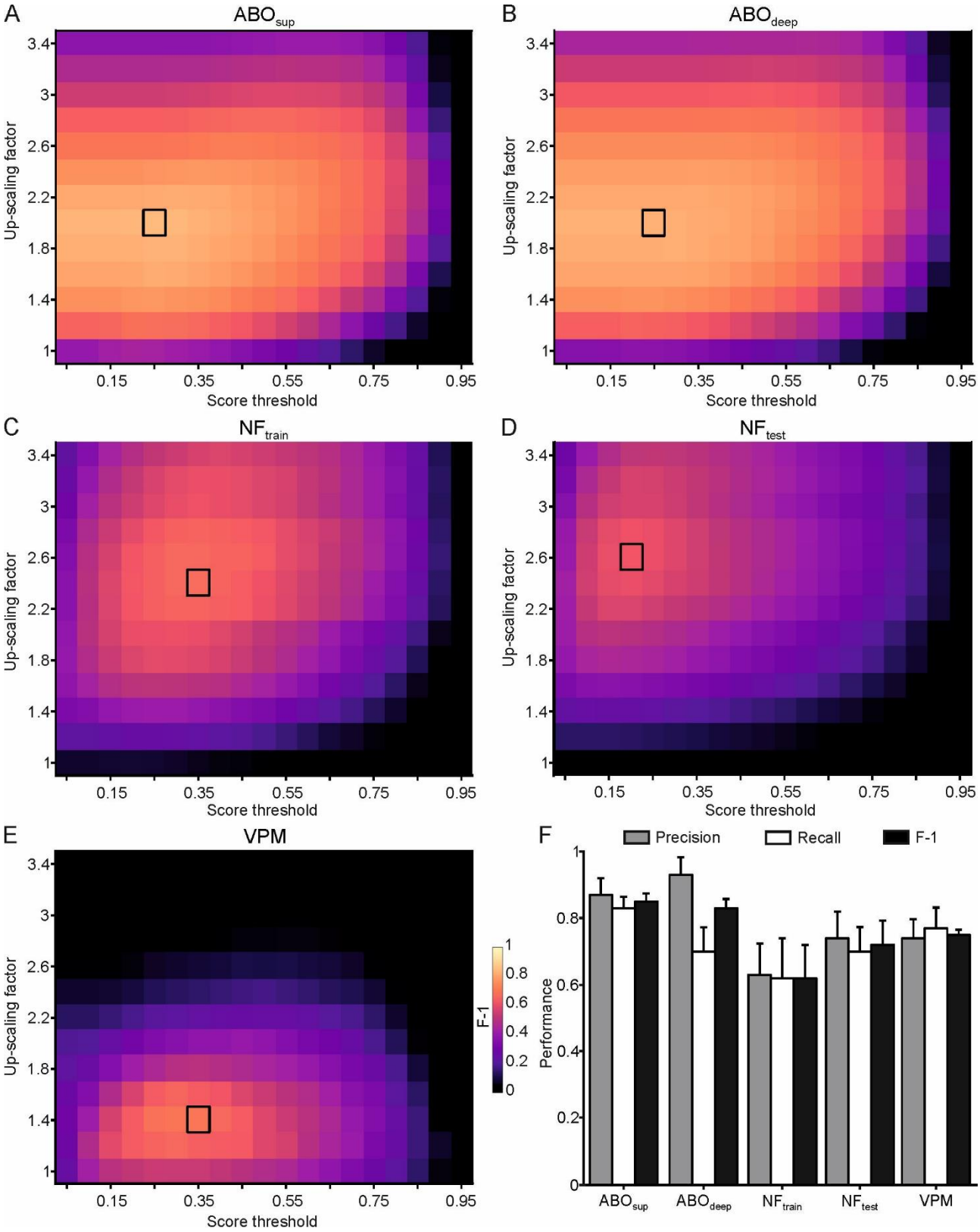


Figure 5. CITE-On offline performance on never-before-seen data. A-E) Best parameter search for frame upscaling: F-1 score (pseudocolor) as a function of upscaling factor (vertical axis) and score

threshold (horizontal axis) for the ABO_{sup} (A), ABO_{deep} (B), NF_{train} (C), NF_{test} (D) and VPM (E) datasets. The maximal F-1 is indicated with the black rectangle. The pseudocolor scale in (E) applies to (A-D). **F)** Performance of Precision (grey), Recall (white), and F-1 (black) for all t-series in the ABO_{sup} (N = 9), ABO_{deep} (N = 10), NF_{train} (N = 19), NF_{test} (N = 9), and VPM (N = 9) datasets.

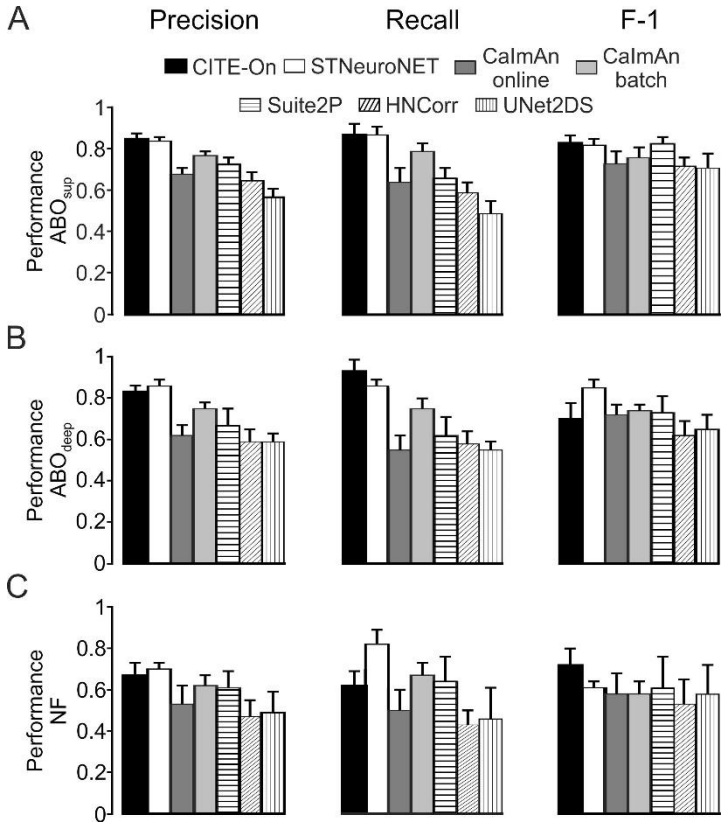


Figure 6. CITE-On offline performance compared with state-of-the-art methods. A-C) Offline performance of CITE-On (black) compared to STNeuroNET (white), CaImAn On Line (dark grey), CaImAn Batch (light grey), Suite2P (horizontal line), HNCorr (tilted line) and UNet2DS (vertical line). Precision (left), Recall (middle), and F-1 (right) are shown. Performance is evaluated on ABO_{sup} (A, N = 9 t-series), ABO_{deep} (B, N = 10 t-series), and NF_{train} + NF_{test} (NF, C, N = 28 t-series).

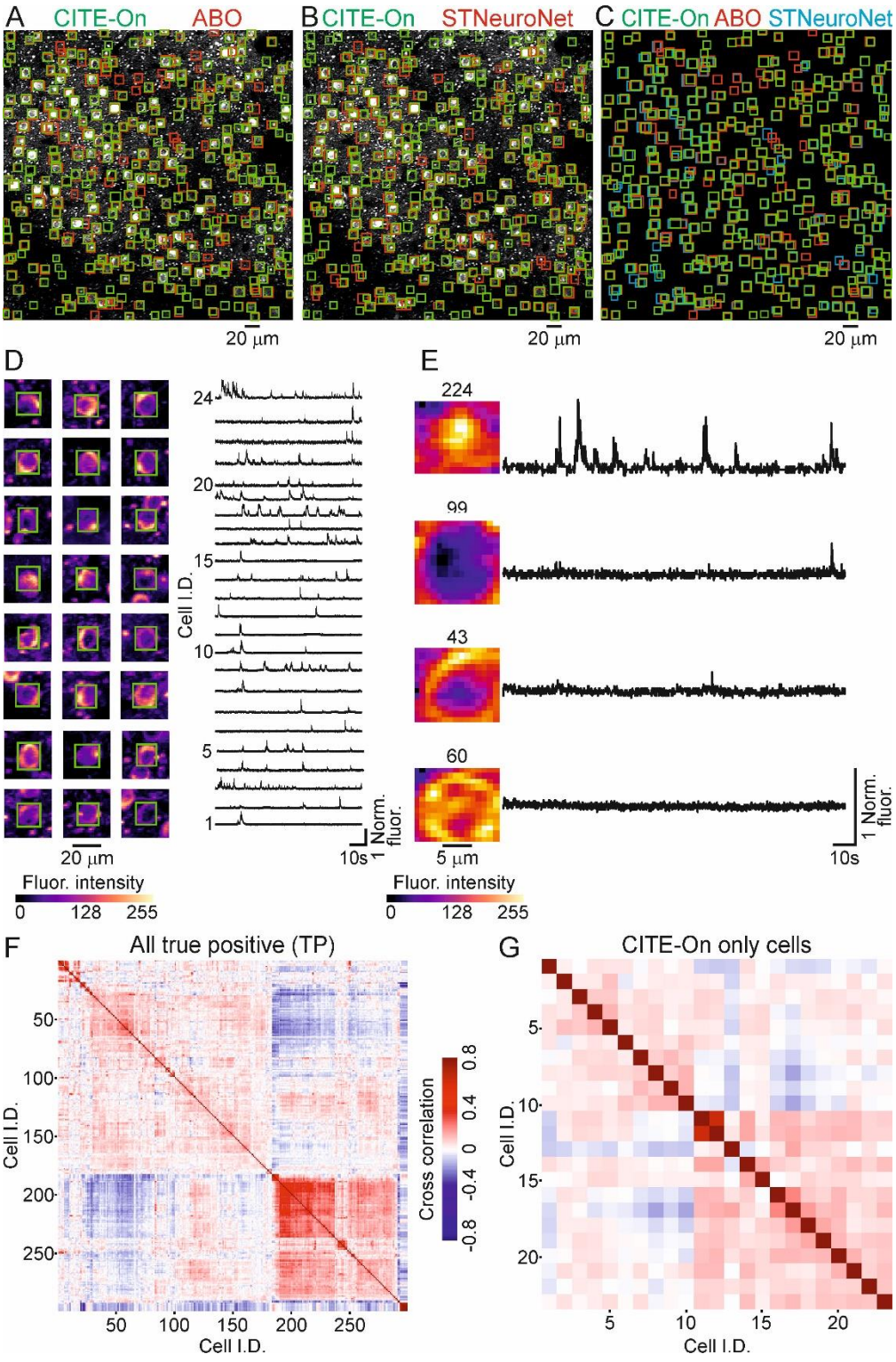


Figure 7. CITE-On data processing of never-before-seen recordings. A) Median projection of a representative t-series from the ABO dataset showing GCaMP6f expressing cortical neurons. CITE-On true positives (CITE-On, green) and true positives provided by the Allen Brain Observatory

(ABO, red) are shown. **B)** Same as in (A) with CITE-On true positives (green) and STNeuroNET true positives (red). **C)** Superposition of CITE-On (green), ABO (red), and STNeuroNET (cyan) true positives. **D)** Left: 24 representative cells detected by CITE-On and identified as true positives in ABO and STNeuroNET. The CITE-On-identified bounding box is represented in green. Right: corresponding CITE-On-extracted fluorescence traces. **E)** Same as in (D) for four representative CITE-On-only cells. These fours cells were not counted in the GT of ABO and STNeuroNET GT, either as true or false positives. **F)** Lower-left triangle: cross correlation matrix for all functional traces extracted from true positive detection in the t-series displayed in (A). Upper-right triangle: corresponding dendrogram sorting. The pseudocolor scale indicates the cross correlation value. **G)** same as in (F), but for the CITE-On-only true positive cells. Pseudocolor scale as in (F).

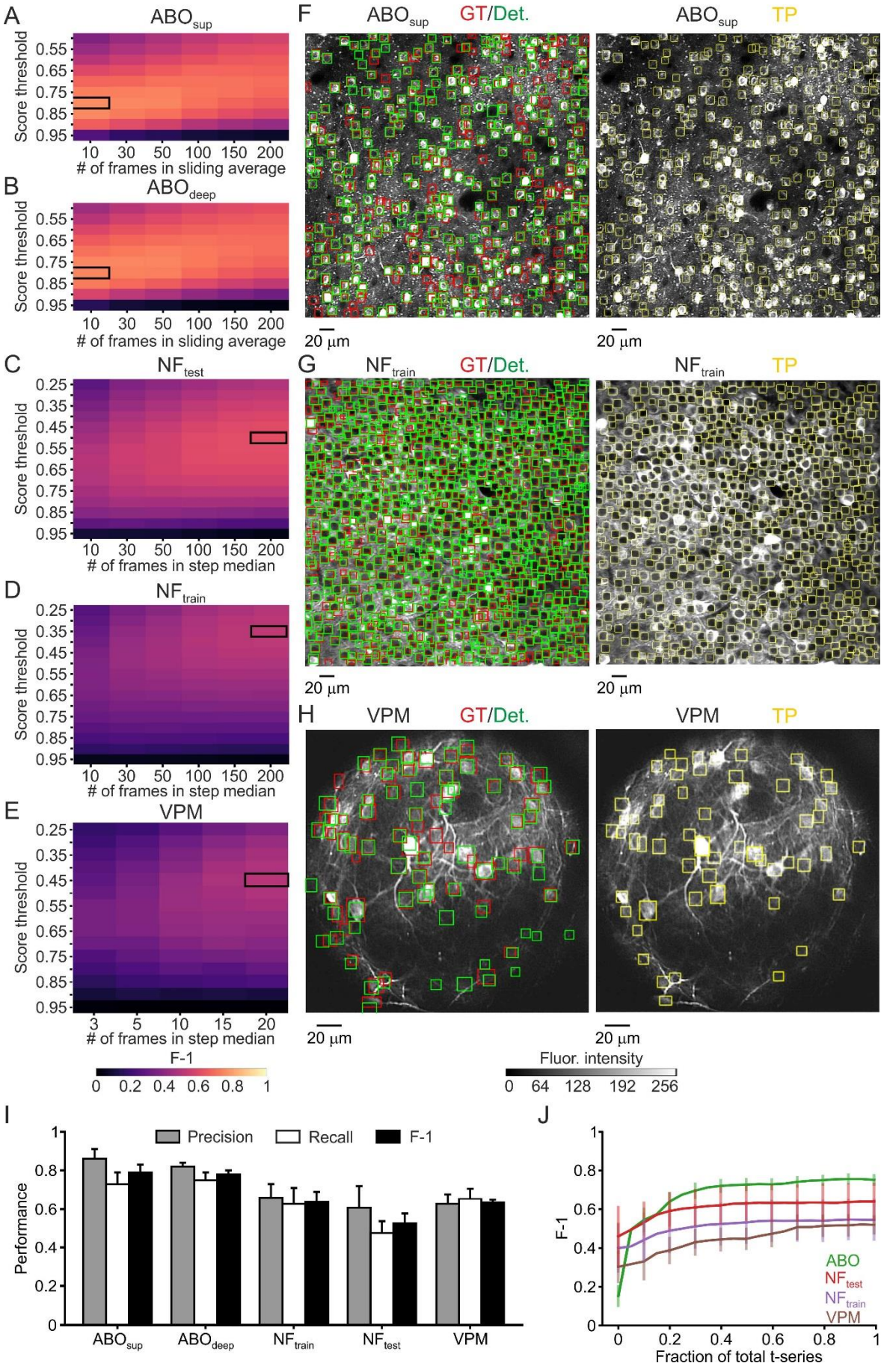

A
ABO sup
Score threshold
0.55
0.65
0.75
0.85
0.95
10 30 50 100 150 200
# of frames in sliding average
B
ABO deep
Score threshold
0.55
0.65
0.75
0.85
0.95
10 30 50 100 150 200
# of frames in sliding average
C
NF test
Score threshold
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95
10 30 50 100 150 200
# of frames in step median
D
NF train
Score threshold
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95
10 30 50 100 150 200
# of frames in step median
E
VPM
Score threshold
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95
3 5 10 15 20
# of frames in step median
F-1
0 0.2 0.4 0.6 0.8 1
F
ABO sup GT/Det.
ABO sup TP
20 µm
20 µm
G
NF train GT/Det.
NF train TP
20 µm
20 µm
H
VPM GT/Det.
VPM TP
20 µm
20 µm
Fluor. intensity
0 64 128 192 256
I
Precision
Recall
F-1
Performance
0
0.2
0.4
0.6
0.8
1
ABO sup
ABO deep
NF train
NF test
VPM
J
F-1
0
0.2
0.4
0.6
0.8
1
ABO
NF test
NF train
VPM
0 0.2 0.4 0.6 0.8 1
Fraction of total t-series

Figure 8. CITE-On online performance on never-before-seen datasets. A-E) Best parameter search for frame downsampling: F-1 score (pseudocolor) as a function of score threshold (vertical axis) and number of frames (horizontal axis) for the ABO_{sup} (A), ABO_{deep} (B), NF_{train} (C), NF_{test} (D) and VPM (E) datasets. The maximal F-1 is indicated with the black rectangle. The pseudocolor scale in (E) applies to (A-D). For the ABO_{sup}, ABO_{deep} datasets the sliding average frame downsampling approach was used, while for the NF_{test}, NF_{train}, and VPM datasets the step average approach was implemented. **F-H)** Left: median projection of a representative t-series from the ABO_{sup} (F), NF_{train} (G), and VPM (H) datasets. GT (red) and online CITE-On detections (green bounding boxes) are shown. Right: bounding boxes (yellow) corresponding to true positives are shown. The greyscale in H applies also to (F-G). **I)** Online detection performance of Precision (grey), Recall (white), and F-1 (black) for all t-series in the ABO_{sup} (N = 9), ABO_{deep} (N = 10), NF_{train} (N = 19), NF_{test} (N = 9), and VPM (N = 9) datasets. **J)** F-1 as a function of the fraction of processed t-series for ABO (green, N = 19 t-series), NF_{test} (red, N = 9 t-series), NF_{train} (purple, N = 19 t-series), and VPM (brown, N = 9 t-series) datasets. 10 frames sliding averages for ABO; detection rate, 5 Hz. Step median of 20 frames and 200 frames for VPM and NF datasets; detection rate, 0.3 Hz and 0.035 Hz for SPM and NF datasets, respectively.

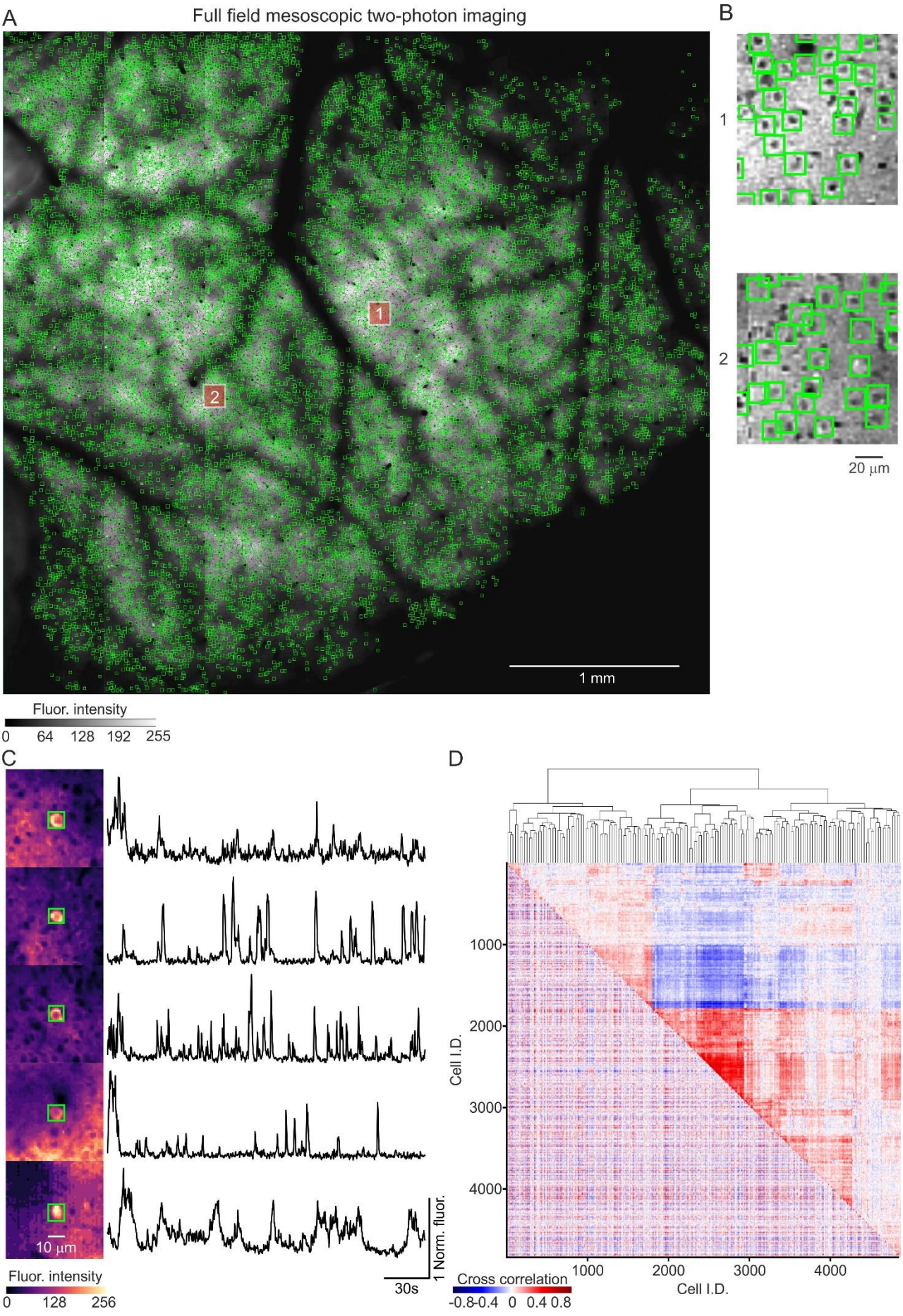
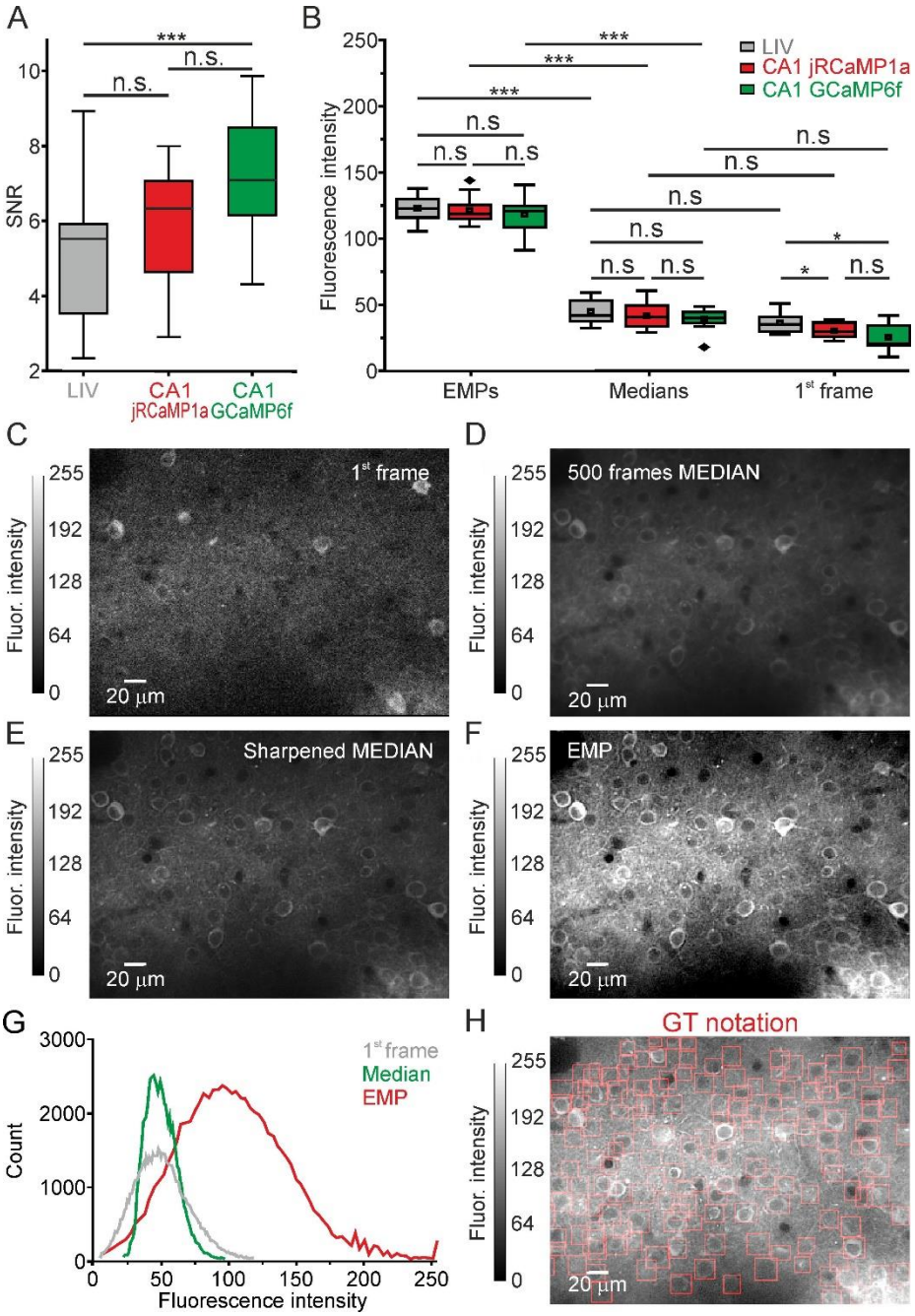

A
Full field mesoscopic two-photon imaging
1
2
1 mm
Fluor. intensity
0 64 128 192 255
B
1
2
20 μm
C
10 μm
Fluor. intensity
0 128 256
30s
1 Norm. fluor.
D
Cell I.D.
1000
2000
3000
4000
Cross correlation
-0.8-0.4 0 0.4 0.8
Cell I.D.

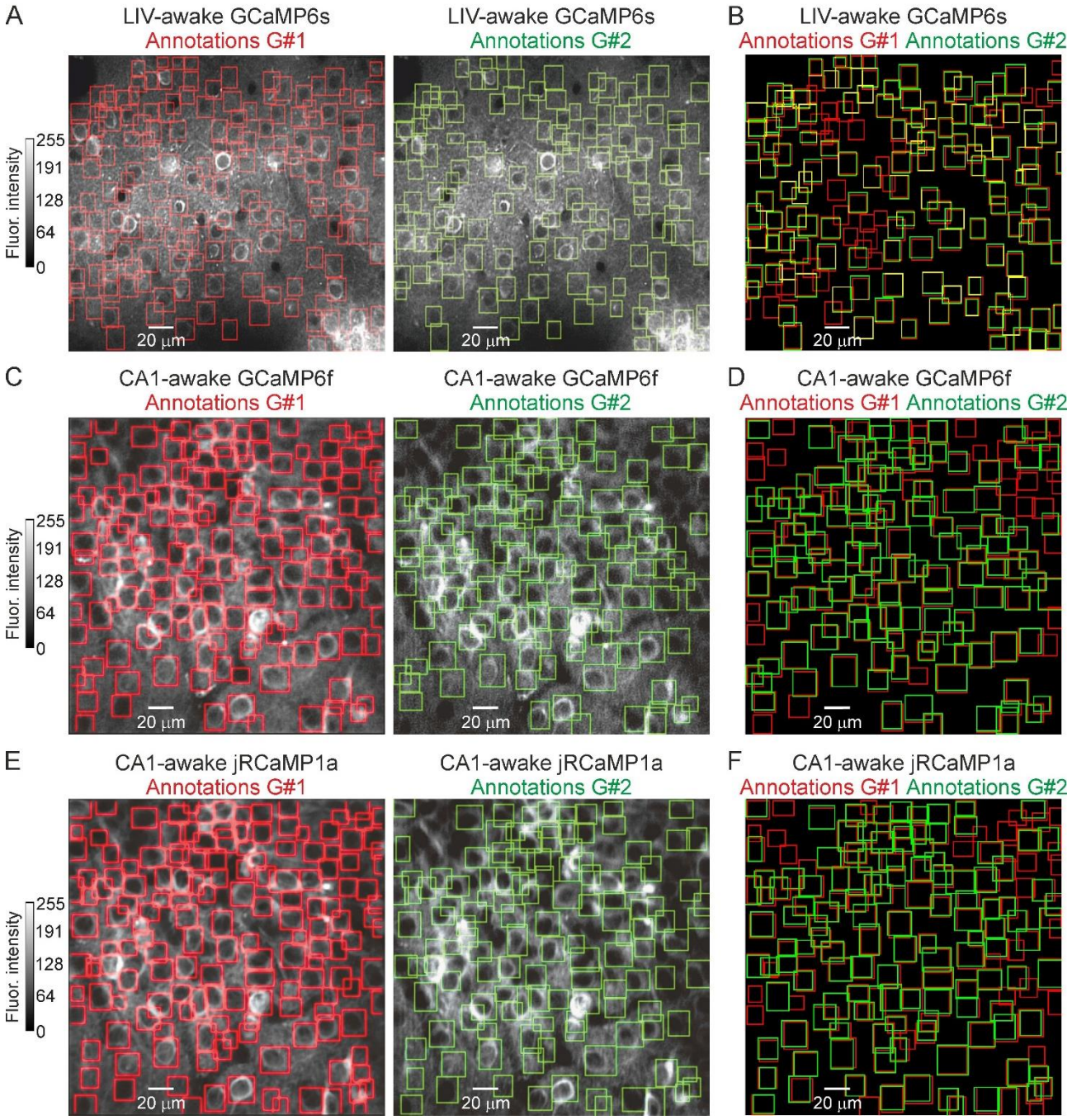
Figure 9. CITE-On analysis of mesoscopic two-photon imaging t-series. A-B) Median projection of a mesocopic imaging t-series showing GCaMP6s expressing neurons (mesoscopic data from [7]). Green boxes indicate cells detected by CITE-On (total: 4,842 cells). Two regions are highlighted by yellow squares and shown at an expanded spatial scale in (B). Greyscale in (A) applies also to (B). **C)** Left: five representative cells detected by CITE-On. Right: corresponding CITE-On-extracted fluorescence traces in the first 230 s of the t-series. **D)** Cross correlation matrix (bottom-left triangle) calculated on the background subtracted traces extracted by CITE-On on all detected cells in the first 7,000 frames and relative dendrogram (top-right triangle).

Supplementary material

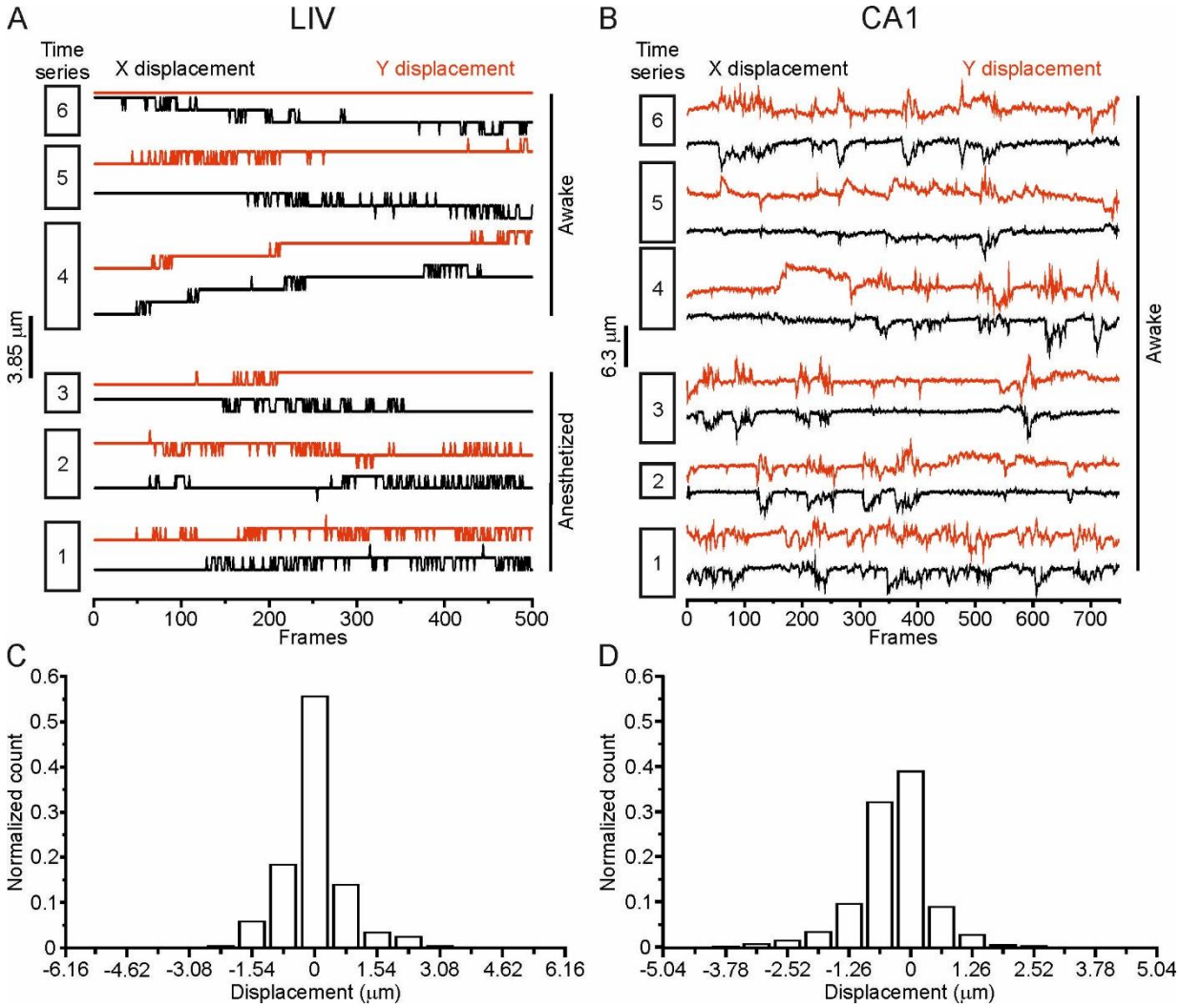


Supplementary Figure 1. Processing imaging t-series for GT annotation. A) SNR values across all dataset t-series (both training and validation). Unpaired Student's *t*-test, p = 0.17 between LIV vs. CA1 jRCaMP1a; p = 0.009 for LIV vs. CA1 GCaMP6f; p = 0.26 for CA1 GCaMP6f vs. CA1 jRCaMP1a. N = 121 for LIV, N = 38 for CA1 jRCaMP1a and for CA1 GCaMP6f. **B)** Fluorescence intensity for EMPs images, median projections of time series, and individual first frames of acquisitions for LIV (grey) and CA1 (GCaMP6f: green and jRCaMP1a: red) datasets. For EMPs, unpaired Student's *t*-test: p = 0.77 for LIV vs. CA1 jRCaMP1a; p = 0.62 for LIV vs. CA1 GCaMP6f; p = 0.78 for CA1 GCaMP6f vs. CA1 jRCaMP1a. For individual frames, Wilcoxon signed rank test: p = 0.38 for LIV vs. CA1 jRCaMP1a; p = 0.56 for LIV vs. CA1 GCaMP6f; p = 0.68 for CA1

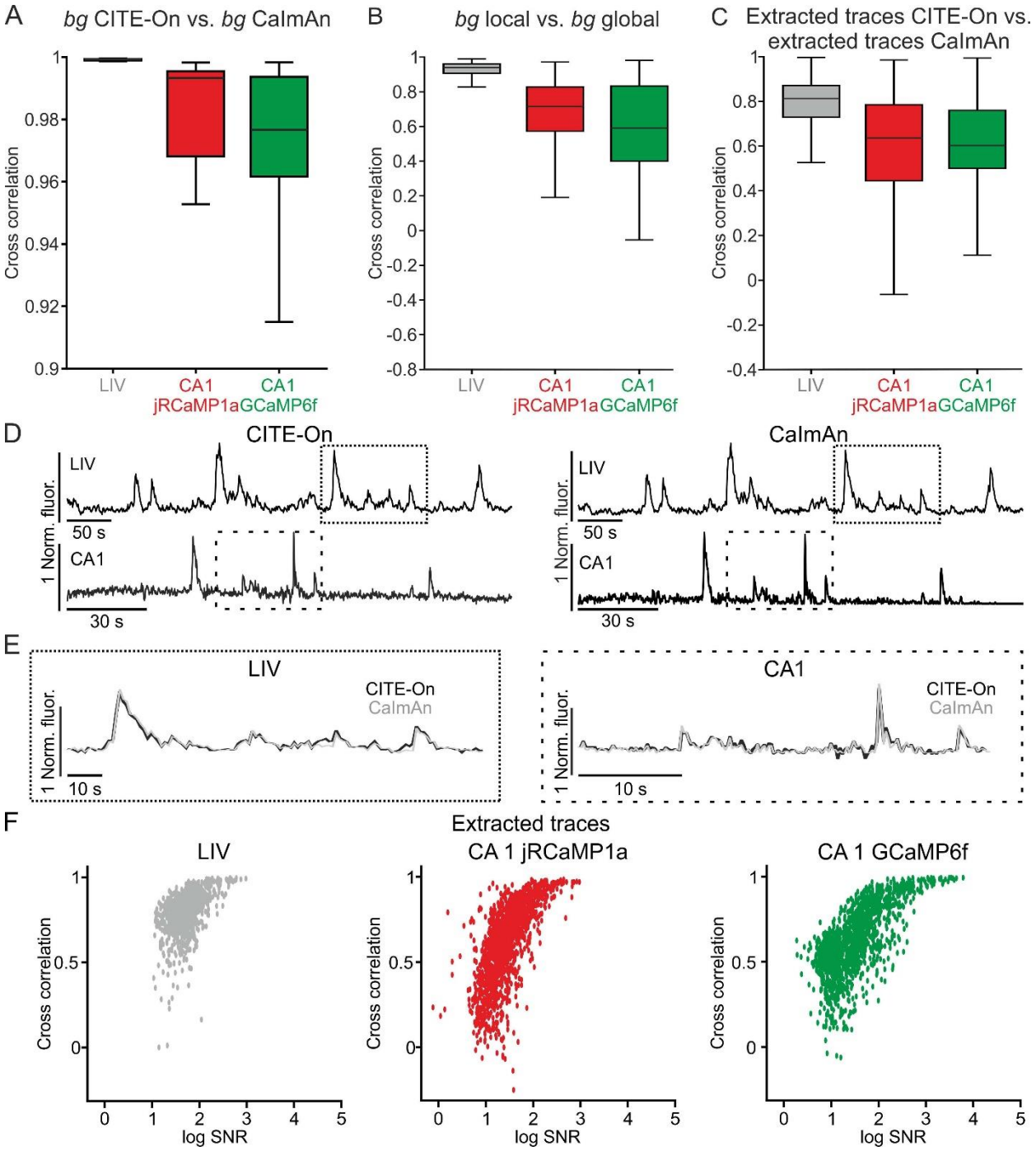
jRCaMP1a vs. CA1 GCaMP6f. For EMPs vs. median values, median values for LIV were significantly larger compared with individual frames of CA1 jRCAMP1a and CA1 GCaMP6f: Wilcoxon signed rank test, p = 2.5E-5 for LIV median vs. 1st frame CA1 jRCaMP1a; p = 8.6E-6 for LIV median vs. 1st frame CA1 GCaMP6f; p = 0.082 for LIV median vs. LIV 1st frame. Wilcoxon signed rank test p = 3.5E-5 for LIV EMP vs. LIV median; p = 4.8E-8 for LIV EMP vs. LIV 1st frame; p = 5.6E-6 for LIV EMP vs. CA1 jRCaMP1a median; p = 4.8E-8 for LIV EMP Vs CA1 jRCaMP1a 1st frame; p = 5.6E-6 for LIV EMP vs. CA1 GCaMP6f median; p = 6.9E-8 for LIV EMP vs. CA1 GCaMP6f 1st frame). N = 121 for LIV, N = 38 for CA1 jRCaMP1a and for CA1 GCaMP6f. **C-F)** Individual frame (C), median projection (D), global sharpened image (E), and EMP (F) for a representative LIV t-series. **G)** Distribution of absolute fluorescence intensity values for individual frames (grey), median projection (green), sharpened median projection (cyan) and EMP images (red) for a representative LIV t-series. **H)** GT annotation on a representative EMP.



Supplementary Figure 2. Graders' annotations on LIV and CA1 validation t-series. A) EMP image from a representative LIV t-series. Bounding boxes generated by grader #1 are shown in red in the left panel. Those generated by grader #2 are shown in green in the right panel. **B)** Superposition of the bounding boxes generated by grader #1 (red) and grader #2 (green). **C-D)** Same as in (A-B) for a representative CA1 GCaMP6f t-series. **E-F)** Same as in (A-B) for a representative CA1 jRCaMP1a t-series.

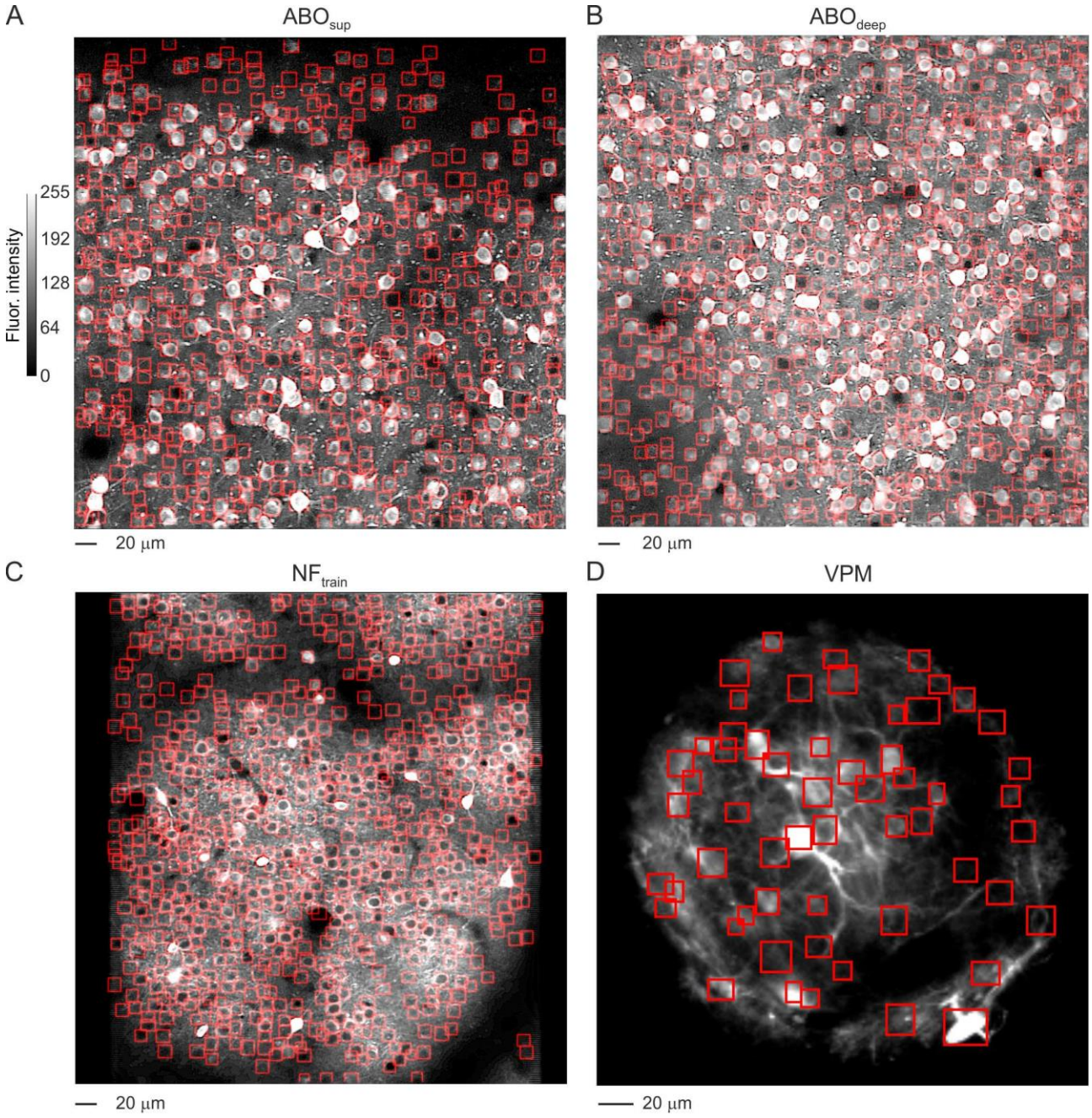


Supplementary Figure 3. **Motion artefacts in the LIV and CA1 validation datasets**. **A)** X, Y displacement of the FOV (black, X; red, Y) expressed in microns as observed across frames of six representative 500 frame-long LIV t-series (traces 1-3 from anesthetized animals, traces 4-6 from awake mice). **B)** Same as in (A) but for six representative 750 frame-long CA1 jRCaMP1a acquisitions. **C)** Percentage of total X, Y displacements in LIV validation t-series (N = 13). D) Same as in (C) for CA1 jRCaMP1a validation t-series (N = 12).

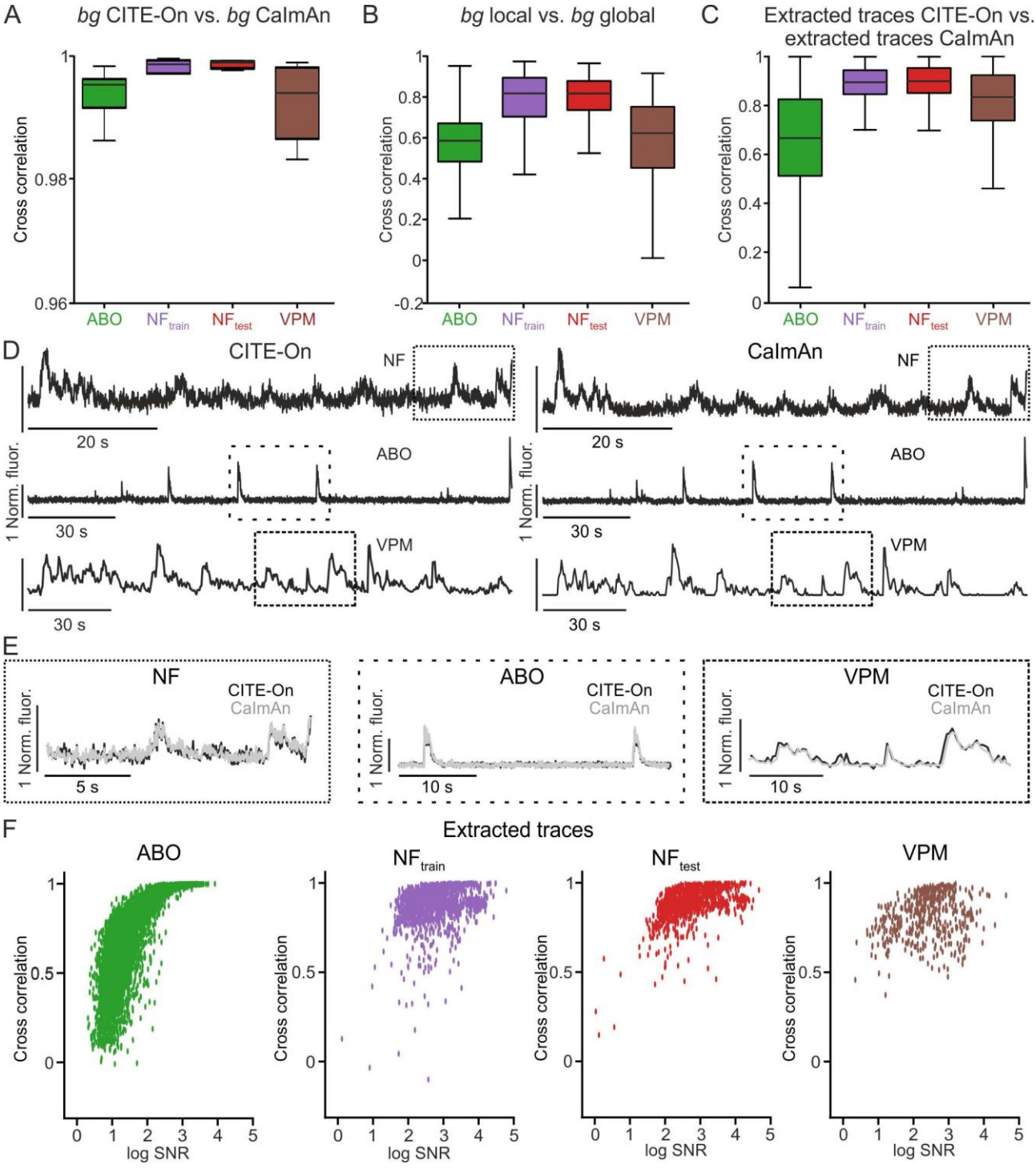


Supplementary Figure 4. Trace extraction from the validation dataset: CITE-On vs. seeded-CaImAn. A) Boxplots showing cross correlation values between the background obtained using CITE-On and seeded-CaImAn for the LIV, CA1 jRCaMP1a, and CA1 GCaMP6f datasets (N = 13, N = 12, and N = 12 t-series, respectively). **B)** Cross correlation values between global and local background signals computed with CITE-On. **C)** Cross correlation of background-subtracted functional traces extracted with CITE-On and with seeded-CaImAn for all true positive detected identities in the LIV, CA1 jRCaMP1a, and CA1 GCaMP6f datasets. **D)** Representative background subtracted functional traces extracted with CITE-On (left) and seeded-CaImAn (right) from LIV (top) and CA1 jRCaMP1a (bottom) t-series. **E)** Representative traces extracted with CITE-On (black) and CaImAn (grey) are shown superimposed for LIV (left) and CA1 (right). **F)** Cross correlation of

background-subtracted functional traces extracted with CITE-On and with seeded-CaImAn as a function of the cell's SNR for all true positive identities in the LIV (left), CA1 jRCaMP1a (middle), and CA1 GCaMP6f (right) datasets. Each dot represents a cell detected by CITE-On (see Supplementary Table 1).

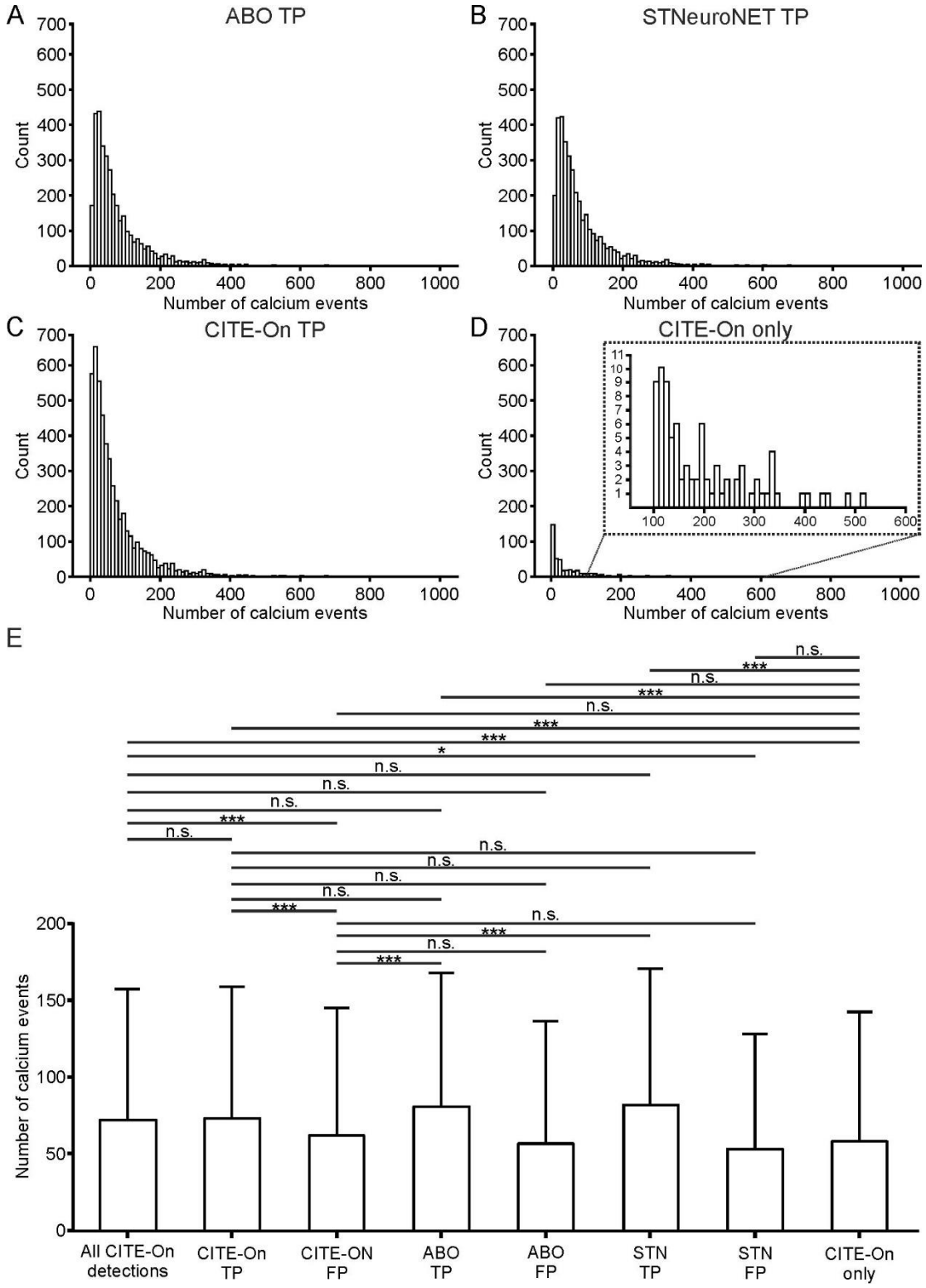


Supplementary Figure 5. GT bounding box annotation of publicly available datasets. A-D) Median projection of t-series from ABO_{sup} (A), ABO_{deep} (B), NF_{train} (C), and VPM (D) datasets. GT bounding boxes are shown in red.



Supplementary Figure 6. Trace extraction from publicly available datasets: CITE-On vs. seeded-CaImAn. A) Boxplots showing cross correlation values between the background obtained using CITE-On and seeded-CaImAn for the for ABO, NF_{train}, NF_{test}, and VPM data (N = 19, N = 19, N = 9, and N = 9 t-series, respectively). **B)** Cross correlation values between global and local background signals computed with CITE-On. **C)** Cross correlation of background-subtracted functional traces extracted with CITE-On and with seeded-CaImAn for the various datasets. **D)** Representative background subtracted functional traces extracted with CITE-On (left) and seeded-CaImAn (right) for NF (top), ABO (middle) and VPM (bottom) acquisitions. **E)** Representative traces

extracted with CITE-On (black) and CaImAn (grey) are shown superimposed for NF (left), ABO (middle), and VPM (right). **F)** Cross correlation of background-subtracted functional traces extracted with seeded-CaImAn and CITE-On as a function of the cell's SNR for all true positive identities in the ABO (leftmost), NF_{train} (middle left), NF_{test} (middle right) and VPM (rightmost) acquisitions. Each dot represents a cell detected by CITE-On (see Supplementary Table 1).



Supplementary Figure 7. Calcium activity in CITE-On, ABO, and STNeuroNET detections. A-D) Distribution of the number of calcium events in true positive detection in a representative ABO t-series analyzed by ABO (A), STNeuroNET (B), CITE-On (C). (D) shows the distribution of CITE-On-only true positive detections. Data refers to the corresponding GT annotation provided by the consensus GT for CITE-On, ABO, and STNeuroNET. **E)** Average (and s.d.) number of detected

calcium events *per* cell for all CITE-On detections, CITE-On true positives (CITE-On TP), CITE-On false positives (CITE-On FP), ABO true positives (ABO TP), ABO false positives (ABO FP), STNeuroNET true positives (STN TP), STNeuroNET false positives (STN FP), and CITE-On only detections. Data from N = 19 t-series of the ABO dataset. Results of Kolmogorov-Smirnov test: p = 0.00012 for All cells vs. CITE-On-only cells; p = 1.00 for All cells vs. CITE-On True Positives; p = 0.00022 for All cells vs. CITE-On False Positives; p = 0.99 for All cells vs. ABO True Positives; p = 0.078 for All cells vs. ABO False Positives; p = 0.99 for All cells vs. STN True Positives; p = 0.036 for All cells vs. STN False Positives; p =0.00041 for CITE-On-only Cells vs. CITE-On True Positives; p = 0.99 for CITE-On-only Cells vs. CITE-On False Positives; p = 0.0012 for CITE-On-only Cells vs. ABO True Positives; p = 0.28 for CITE-On-only Cells vs. ABO False Positives; p = 0.00073 for CITE-On-only Cells vs. STN True Positives; p = 0.36 for CITE-On-only Cells vs. STN False Positives; p = 0.00073 for CITE-On True Positives vs. CITE-On False Positives; p = 0.99 for CITE-On True Positives vs. ABO True Positives; p = 0.11 for CITE-On True Positives vs. ABO False Positives; p = 0.99 for CITE-On True Positives vs. STN True Positives; p = 0.078 for CITE-On True Positives vs. STN False Positives; p = 0.0022 for CITE-On False Positives vs. ABO True Positives; p = 0.46 for CITE-On False Positives vs. ABO False Positives; p = 0.0012 for CITE-On False Positives vs. STN True Positives; p = 0.70 for CITE-On False Positives vs. STN False Positives; p =0.28 for ABO True Positives vs. ABO False Positives; p = 1.00 for ABO True Positives vs. STN True Positives; p = 0.15 for ABO True Positives vs. STN False Positives; p = 0.21 for ABO False Positives vs. STN True Positives; p = 0.99 for ABO False Positives vs. STN False Positives; p =0.11 for STN True Positives vs. STN False Positives. Total number of detected cells: N = 5482, for All CITE-On detections; N = 4934, for CITE-On TP; N = 548, for CITE-On FP; N = 3516, for ABO TP; N = 1966, for ABO FP; N = 3606, for STN TP; N = 1876, for STN FP; N = 439, for CITE-On only.

Supplementary tables

Dataset	# of t-series	# of detections grader 1	# of detections grader 2	# of detections overlap	# of detections grader 1 only	# of detections grader 2 only	# of ground truth
ABO	19	6225	5456	5443	782	13	6238
NF_{train}	19	11171	9772	9630	1541	142	11313
NF_{test}	9	3187	2814	2793	394	21	3208
VPM	9	517	470	444	73	26	543
LIV_{test}	13	979	868	861	118	7	986
$CA1_{test}$ jRCaMP1a	12	884	809	795	89	14	898
$CA1_{test}$ GCaMP6s	12	1000	927	916	84	11	1011
LIV_{train}	118	7221	6489	6356	865	133	7354
$CA1_{train}$ jRCaMP1a	21	2123	1856	1833	290	23	2146
$CA1_{train}$ GCaMP6s	21	2218	1963	1940	278	23	2241

Supplementary table 1. Dataset annotation and ground truth generation. Two graders manually annotated the LIV, CA1, ABO, NF, and VPM datasets. The table reports the number of t-series, the

number of detections by grader 1 and 2, the number of overlapping detections, the number of detections exclusively produced by grader 1 and grader 2, and the number of detections in the consensus ground truth.

Dataset	mAP	s.d	N	F-1	s.d	N	Precision	s.d	N	Recall	s.d	N
ABO	0.80	0.10	19	0.93	0.02	19	0.998	0.004	19	0.87	0.04	19
NF_{train}	0.71	0.14	19	0.92	0.03	19	0.99	0.02	19	0.86	0.05	19
NF_{test}	0.72	0.10	9	0.93	0.02	9	0.99	0.01	9	0.88	0.04	9
VPM	0.64	0.13	9	0.90	0.03	9	0.95	0.03	9	0.86	0.04	9
LIV_{test}	0.81	0.11	13	0.93	0.03	13	0.99	0.01	13	0.88	0.05	13
$CA1_{test}$ jRCaMP1a	0.76	0.07	12	0.94	0.03	12	0.98	0.02	12	0.90	0.04	12
$CA1_{test}$ GCaMP6s	0.82	0.08	12	0.95	0.01	12	0.99	0.01	12	0.92	0.03	12
LIV_{train}	0.73	0.09	118	0.93	0.02	118	0.98	0.02	118	0.88	0.05	118
$CA1_{train}$ jRCaMP1a	0.74	0.08	21	0.92	0.02	21	0.99	0.01	21	0.87	0.04	21
$CA1_{train}$ GCaMP6s	0.76	0.07	21	0.93	0.02	21	0.99	0.01	21	0.87	0.04	21

Supplementary table 2. Evaluation of consensus GT. The table reports the average (and relative s.d.) of the mean average precision (mAP), the F-1 score, the precision and recall across the t-series of the various datasets annotated by the different graders.

Offline												
Dataset	F-1	s.d	N	Precision	s.d	N	Recall	s.d	N	# of TP	# of FP	# of FN
ABO	0.84	0.03	19	0.90	0.06	19	0.80	0.07	19	4934	548	1303
NF_{train}	0.63	0.1	19	0.63	0.09	19	0.62	0.12	19	6849	3932	4120
NF_{test}	0.74	0.08	9	0.70	0.07	9	0.51	0.06	9	2286	716	911
VPM	0.75	0.02	9	0.74	0.06	9	0.77	0.06	9	398	140	123
LIV_{test}	0.86	0.04	13	0.85	0.05	13	0.88	0.05	13	1486	257	205
$CA1_{test}$ jRCaMP1a	0.92	0.04	12	0.88	0.06	12	0.97	0.02	12	1618	222	55
$CA1_{test}$ GCaMP6s	0.93	0.02	12	0.91	0.03	12	0.97	0.02	12	1638	170	56

Supplementary table 3. CITE-On offline performance. Average (and relative s.d.) F-1, Precision, Recall, and number of true positives (TP), false positives (FP), and false negatives (FN) for LIV, CA1, ABO, NF and VPM datasets generated by the CITE-On offline pipeline.

Online

Dataset	F-1	s.d	N	Precision	s.d	N	Recall	s.d	N	# of TP	# of FP	# of FN
ABO	0.77	0.03	19	0.83	0.04	19	0.83	0.04	19	4573	978	1664
NF_{train}	0.55	0.11	19	0.56	0.11	19	0.56	0.11	19	6164	4906	5026
NF_{test}	0.64	0.10	9	0.68	0.12	9	0.68	0.12	9	2096	921	1101
VPM	0.52	0.50	9	0.60	0.11	9	0.60	0.11	9	244	181	277
LIV_{test}	0.69	0.04	13	0.72	0.05	13	0.72	0.05	13	1127	437	564
$CA1_{test}$ jRCaMP1a	0.73	0.07	12	0.74	0.07	12	0.74	0.07	12	1233	439	440
$CA1_{test}$ GCaMP6s	0.67	0.12	12	0.69	0.12	12	0.69	0.12	12	1145	493	549

Supplementary table 4. CITE-On online performance. Average (and relative s.d.) F-1, Precision, Recall, and number of true positives (TP), false positives (FP), and false negatives (FN) for LIV, CA1, ABO, NF and VPM datasets generated by the CITE-On online pipeline.

ABO Dataset	# of detections	s.d	N
All CITE-On	289	21	19
CITE-On TP	260	28	19
CITE-On FP	29	16	19
ABO TP	185	36	19
ABO FP	103	31	19
STNeuroNET TP	190	33	19
STNeuroNET FP	99	24	19
CITE-On only	23	15	19

Supplementary table 5. Detections in the ABO dataset. Average (and relative s.d.) number of detections in the ABO dataset reported by CITE-On, ABO, and STNeuroNET. TP, true positives; FP, false positives. Data relative to CITE-On were calculated online as the number of identities obtained at the end of the processing of each t-series then averaged across all t-series.

Supplementary movie legends

Supplementary movie 1. Online tracking of detected identities across the t-series. A representative t-series (# of frames, 750; frame rate, 3 Hz; no motion correction) from the CA1 validation dataset was processed online for cell detection (detection update rate, 10 Hz). Bounding boxes (colored squares) for active detections (i.e. neurons identified by CITE-On in the current frame) and past detections (i.e. neurons identified by CITE-On in previous frames) are represented with or without a central dot, respectively. Each identity was associated with a bounding box color which was retained across the t-series. The position and shape of each bounding box was updated in each frame according to the procedure described in the Results section.

Supplementary movie 2. Frame-by-frame manual annotation. Representative t-series (# of frames, 500; frame rate, 1.5 Hz) from the LIV training dataset (awake head-fixed *Scnn1a-cre* mouse expressing GCaMP6s) used for manual annotation of individual frames. The t-series was motion corrected. Brightness and contrast of each displayed frame of the t-series could be adjusted. Bounding boxes were manually positioned around each visible cell in each frame. Only a minority of neurons were identified in the frame-by-frame annotation process.

Supplementary movie 3. Manual annotation of EMP images. EMP image corresponding to the same t-series as in Supplementary movie 2. Brightness and contrast of the EMP image could be regulated to better visualize bright and dim cells. Bounding boxes (green) were manually defined around identified neurons.

Supplementary movie 4. Online dynamic segmentation and functional trace extraction. Top left: raw fluorescence of a representative cell from a t-series in the ABO dataset. The displayed cell was detected online using CITE-On. Top right: binary mask generated by CITE-On on each frame of the t-series displayed on the left. The white pixels represent the segmented pixels, black pixels are discarded. Bottom: functional fluorescence trace extracted online for the cell displayed on top.